Simultaneous Three-Dimensional Mapping and Geolocation of Road Surface

Diya Li

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
in
Mechanical Engineering

Tomonari Furukawa, Chair
Corina Sandu
Kevin Kochersberger

September 17, 2018
Blacksburg, Virginia

Keywords: Extended Kalman Filter, Global Localization, Adaptive Estimation, Road Surface Mapping, Sparse Global Position, Visual Odometry

Simultaneous Three-Dimensional Mapping and Geolocation of Road Surface

Diya Li

ACADEMIC ABSTRACT

This thesis paper presents a simultaneous 3D mapping and geolocation of road surface technique that combines local road surface mapping and global camera localization. The local road surface is generated by structure from motion (SFM) with multiple views and optimized by Bundle Adjustment (BA). A system is developed for the global reconstruction of 3D road surface. Using the system, the proposed technique globally reconstructs 3D road surface by estimating the global camera pose using the Adaptive Extended Kalman Filter (AEKF) and integrates it with local road surface reconstruction techniques. The proposed AEKF-based technique uses image shift as prior. And the camera pose was corrected with the sparse low-accuracy Global Positioning System (GPS) data and digital elevation map (DEM). The AEKF adaptively updates the covariance of uncertainties such that the estimation works well in environment with varying uncertainties. The image capturing system is designed with the camera frame rate being dynamically controlled by vehicle speed read from on-board diagnostics (OBD) for capturing continuous data and helping to remove the effects of moving vehicle shadow from the images with a Random Sample and Consensus (RANSAC) algorithm. The proposed technique is tested in both simulation and field experiment, and compared with similar previous work. The results show that the proposed technique achieves better accuracy than conventional Extended Kalman Filter (EKF) method and achieves smaller translation error than other similar other works.

Simultaneous Three-Dimensional Mapping and Geolocation of Road Surface

Diya Li

GENERAL AUDIENCE ABSTRACT

This thesis paper presents a simultaneous three dimensional (3D) mapping and geolocation of road surface technique that combines local road surface mapping and global camera localization. The local road surface is reconstructed by image processing technique with optimization. And the designed system globally reconstructs 3D road surface by estimating the global camera poses using the proposed Adaptive Extended Kalman Filter (AEKF)-based method and integrates with local road surface reconstructing technique. The camera pose uses image shift as prior, and is corrected with the sparse low-accuracy Global Positioning System (GPS) data and digital elevation map (DEM). The final 3D road surface map with geolocation is generated by combining both local road surface mapping and global localization results. The proposed technique is tested in both simulation and field experiment, and compared with similar previous work. The results show that the proposed technique achieves better accuracy than conventional Extended Kalman Filter (EKF) method and achieves smaller translation error than other similar other works.

Dedication

This thesis is dedicated to my loving parents, Xiaolin Li and Liping Qian.

For their endless love, support and encouragement

Acknowledgments

First and foremost, I would like to sincerely thank my academic advisor, Prof. Tomonari Furukawa, for his continuous guidance and support throughout this study and paper writing. I am also indebted to Prof. Corina Sandu and Prof. Kevin Kochersberger for their services on my advisory committee. Their advice were very beneficial in my completion of the manuscript. I express my heartfelt gratefulness for all committee's guide and support that I believed I learned from the best.

I would like to thank my colleague, Yazhe and Josiah, for their great contribution during experiments. Also, I would like to thank all the people in the Computational Multi-physics Systems Laboratory for their help and inspiration in research and life. Your friendship makes my life in graduate school a wonderful experience. I can not list all the names here, but you are always in my mind.

I would like to take this opportunity to say warm thanks to Yufeng, Yun, Prashant, Dr. Rice and all my beloved friends, who have been so supportive along the way of doing my thesis. I would not be in this position without their help and encouragement.

I must express my wholehearted gratitude to my family for their generous love and support throughout my life and the years of study. This accomplishment would not have been possible without them. I love you all beyond words.

Last but not least, deepest thanks to all people who took part in making this thesis real.

Contents

List of Figures

List of Tables

List of Abbreviations

3D Three Dimensional

AEKF Adaptive Extended Kalman Filter

B-spline Basis spline

EKF Extended Kalman Filter

FOV Field of View

FPGA Field Programmable Gate Array

GNSS Global Navigation Satellite Systems

GPS Global Position System

KF Kalman Filter

MLS Mobile Laser Scanning

MMS Mobile Mapping Systems

OBD On-Board Diagnostics

RANSAC Random Sampling and Consensus

Chapter 1

Introduction

Driving is a common form of transportation in our lives. From the estimates by the Bureau of Transportation Statistics, vehicles dominate the passenger transportation with over 80% [1] which includes cars, trucks, vans, motorcycles, etc. The road surface plays a significant role during driving, which has directly contact with vehicle tires. Therefore, acquiring rich road surface information can not only help civil engineers achieve effective road condition assessment, management and maintenance, but also improve the driving safety and driver comfort. In addition, recent years have seen the growing need of an up-to-date database of existing public road surface, especially in autonomous driving communities [27].

Road surface mapping is the most common technique in collecting road surface information. Previous work indicates that laser scanner and camera are the sensors used and they have involved large manual work. Therefore, many recent researchers developed various automated road information collection system with mobile vehicle. Whereas vision-based technique can extract rich road surface information, the work described here uses vision-based technique and details simultaneous 3D mapping and geolocation of road surface which combines both local road surface mapping techniques and camera global localization. This section provides a brief explanation of background and followed by the objectives of this thesis. The summary of contributions will be listed as well as the general outline of the thesis.

1.1 Background and Motivation

The need of developing accurate high resolution road surface map and more location-based services are emerging not only for the civil infrastructural maintenance, but also for the future autonomous driving that needs to see the road surface condition in real time. Building up a 3D road surface map is an efficient way to inspect road geometry and road surface condition, which can extend the life of the road, reduce the cost of road maintenance and rehabilitation cost, and improve drivers safety and comfort [2, 3].

The traditional method of road surface mapping is using stationary terrestrial laser scanning (STLS) which performs at a static vantage point to collect the road surface point clouds. Recent studies investigated integrated mobile road surface monitor system which not only increases the efficiency of work, but also reduce the risk of human worker, especially in danger environments like highway. One of the common method is laser-based Mobile Mapping Systems (MMS) [26, 39, 51] which is more efficient than STLS to obtain 3D road texture information, but it needs high accurate position sensors for orienting all generated point clouds in world coordinate. The other popular method is image-based method [8, 9, 20] which can provide more details of road surface texture. With orienting all road surface images in world coordinate, the images can be stitched for generating a map.

From the road map shown in the above literatures, most of them only show the results in short segments to analysis in marco-texture level. However, having long road surface map with geolocation will provide more road surface information, such as geometry and it will be much more efficient and reliable. As most of systems are limited in collecting data in long distance, the main motivation of this thesis is to generate road surface map in world coordinate with determining the camera pose using monocular vision algorithm and recursive Bayesian techniques to reduce the translation error in long range, and combing

with vision-based road surface reconstruction technique.

To efficiently represent road surface information in global scale, such as road geometry, road environment and corresponding locations, the data acquisition system is designed and the system that builds up a high resolution 3D public road surface map in world coordinate by integrating multiple sources for road surface mapping is proposed.

1.2 Aim and Objectives

Automatic monitoring short segments of road surface and 3D global localization in long distance has been researched for several decades, but the global road surface mapping and the integration of both technique for building 3D road surface map have not been reached so far. Therefore, to build up 3D road surface map with geolocation, the primary objectives of this thesis are the following:

1. Develop a dynamic frame rate control system for efficiently collecting continuous road surface image along the route.

2. Develop a high resolution road surface mapping technique with geolocation which reduces the scale drift and cumulated position error from vision.

3. Validate the proposed technique and generate the road surface 3D map in global with optimized image poses.

1.3 Original Contribution

The automated road surface surveying has been researched for a long time, but the system for long distance road surface mapping has not reached the maturity. Therefore, the main contributions of this thesis are multi-sensor data acquisition, global camera localization and global road surface reconstruction and visualization, which are summarized as follow:

1. Development of the dynamic frame rate control system along with the road surface image capturing system. The distance between two neighboring images is close to constant which helps for optimizing consecutive image transformation and final road surface patches stitching.

2. Development of the fusion algorithm which optimize the camera global pose with image information, GPS readings, 3D road terrains map as well as the adaptive extended Kalman filter (AEKF).

3. Development of 3D global road surface map with geolocation with local road surface reconstruction and camera global localization techniques and road surface map visualization in world coordinate

1.4 Publication

- **D. Li**, Y. Hu, and T. Furukawa, AEKF-Based 3-D Localization of Road Surface Images with Sparse Low-Accuracy GPS Data, *IEEE 88th Vehicular Technology Conference (VTC)*, August 26-30, 2018, Chicago, USA

- T. Furukawa, **D. Li** and Y. Hu, "High-Resolution Terrain and Road Mapping Using Adaptive Extended Kalman Filter and Enhanced Structure from Motion," *2018 JASE*

Congress, October 17-19, 2018, Nagoya, Japan (In print)

- **D. Li** and T. Furukawa, "Global Vision-Based Reconstruction of Three-Dimensional Road Surfaces Using Adaptive Extended Kalman Filter," *IEEE International Conference on Robotics and Automation (ICRA)*, May 20-24, 2019, Montreal, Canada (Submitted)

- **D. Li** and T. Furukawa, High Resolution Global 3D Road Surface Mapping with Integrating Multi-Sensor, *The Journal of Dynamic Systems, Measurement, and Control, ASME*, (In preparation)

1.5 Outlines

The organization of this thesis are list as the following:

- Ch. 2 presents a comprehensive literature review of road surface mapping techniques
- Ch. 3 explains the fundamental of road surface mapping and the existed problem in the present method
- Ch. 4 explains the proposed AEKF-based camera global localization techniques with camera pose prediction and correction stages
- Ch. 5 explains the proposed global road surface mapping technique with combing the optimized camera pose
- Ch. 6 presents the hardware architecture of proposed system, and discusses the simulation and field experiment results
- Ch. 7 presents the conclusion and future work of this thesis

1.6 Summary

This chapter gives the brief introduction of the road surface mapping, the limitations of current techniques and challenges of using limited source of sensors. This thesis demonstrates how the proposed simultaneous 3D mapping and geolocation of road surface technique combining the local road surface mapping and camera global localization techniques. The road surface images are used for both localization and road surface reconstruction. The interpolation method with B-spline curve is proposed to solve the sparse GPS data with the proposed AEKF model that improves the robustness of localization accuracy in dynamic environment. And the final road surface map is generated with reconstructed road surface segments and the optimized camera poses.

Chapter 2

Review of Literature

This chapter provides the frame of references for this thesis, includes the background and significant technologies survey of road surface mapping and vision-based global localization. The chapter starts with the overview of existed technologies of road surface mapping system in Sec. 2.1. Then Sec. 2.2 reviews computer vision technology applied in localization and some works applied in road surface mapping. The main classifications of literature reviews are shown in Fig. 2.1.

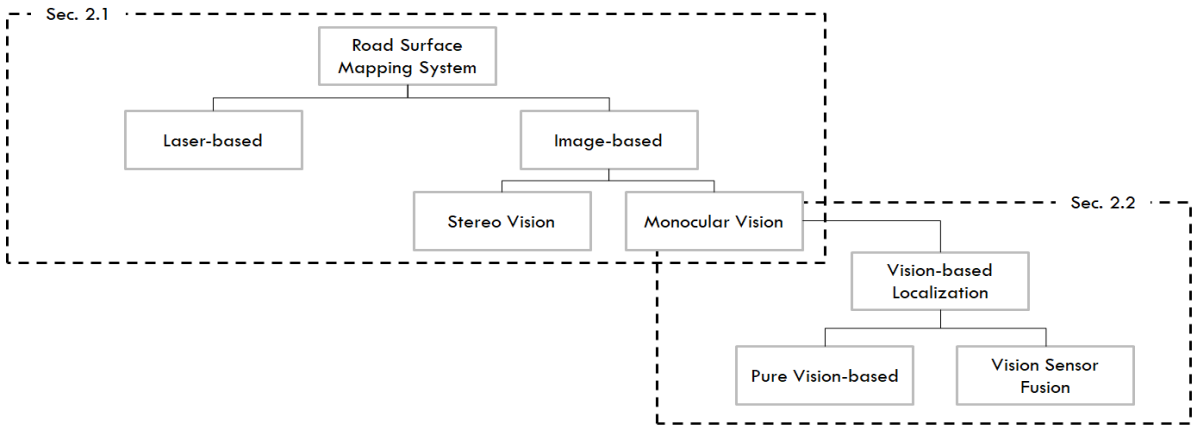


Figure 2.1: The frame of literature reviews

2.1 Mobile Road Surface Mapping System

Human observation is the most common used method in the inspection of road surface; however, this is extremely dangerous and labor intensive, especially in highway. Therefore, an efficient automatic road surface mapping system is needed and has been explored in cost-effective way. The road surface mapping method can be classified into two main categories, laser-based and image-based, based on the sensor used for data collection. This section will review the existing technologies of automatic road surface mapping system.

2.1.1 Laser-Based Road Surface Mapping

Currently, there are various methods for acquiring large-scale 3D laser scan data, such as Mobile Laser Scanning (MLS), Airborne Laser Scanning and Terrestrial Laser Scanning. MLS is useful to acquire dense point clouds of road surface in both industrial and academia. Laser-based road surface mapping system is known as line-based method which is established based on the basis of laser data collected by a MLS. The system consists of 2D laser scanners that measure the distance between road surface and the sensor, and position sensors for data registration, such as GPS and odometer. As shown in Fig. 2.2, the system scans one line at a time and compile the captured point clouds together to form a 3D surface.

Some researchers applied laser-based road surface mapping system in short segment and flat surface, therefore their designed systems do not include position sensor or only include the inertia sensor or odometer to improve the point cloud registration. Kumar et al. [31] designed a laser-based road surface mapping system with two laser scanners and operating simultaneously at fixed rate. To overcome the registration problem, the system is limited to move at the constant speed or fixed line intervals [39]. Jaakkola et al. [26] used a laser scanner and an accurate Inertial Navigation System (INS) for enhanced road surface

Figure 2.2: Laser-based road surface mapping system

mapping with position and attitude information for each scan. The system synchronizes the position and data logging system and retrieves the road surface including paintings and kerbstones by the scanned point cloud data.

Besides INS, the other common position sensor integrated with the laser scanner system is GPS. Yang et al. [51] employed a moving window filtering operation on the mobile laser scanning (MLS) data to classify the road surface from 3D point cloud and further used the GPS times to partition the points into different road sections. As the position of laser scanning data is highly rely on external sensors, most of MLS system used high accuracy position sensors, such as the integration of Global Navigation Satellite System (GNSS) and Inertial Measurement Unit (IMU) [19, 25, 53]. This type of system uses GPS time as the main time system for both laser scanner and IMU to acquire the position and attitude in world coordinate for each scan.

Laser-based road surface mapping technique uses 2D laser scanners integrated with position sensors to produce 3D road surface map. The mobile road surface mapping system provides high accuracy and precision in distance measurement and generates 3D models of road surface. The biggest shortcoming of this technique is that it is hard to find connections

between two 2D laser scanning point clouds so that the system has to be integrated with external position sensor to solve the irregular jump between the two adjacent scan lines. Most of systems employs GPS, however, as on the ground, satellites signal may be obscured so that the positional accuracy will be vary.

2.1.2 Image-Based Road Surface Mapping

Image-based method is an area-based method, which can scan a 2D area at a time as shown in Fig. 2.3. The system can include one or multiple cameras to cover the interested region. From the collected images, the system can provide descriptive information of road surface by image processing. Also, based on the geometry of cameras, the system can also derive the geometric information from 3D point clouds of road surface. The most common image sensor used in image-based road surface mapping is charged Couple Device (CCD) which is a light-sensitive silicon chip. Recently years, Complementary Metal Oxide Semiconductor (CMOS) chip is becoming more popular as it is less energy consuming and performs better [48].

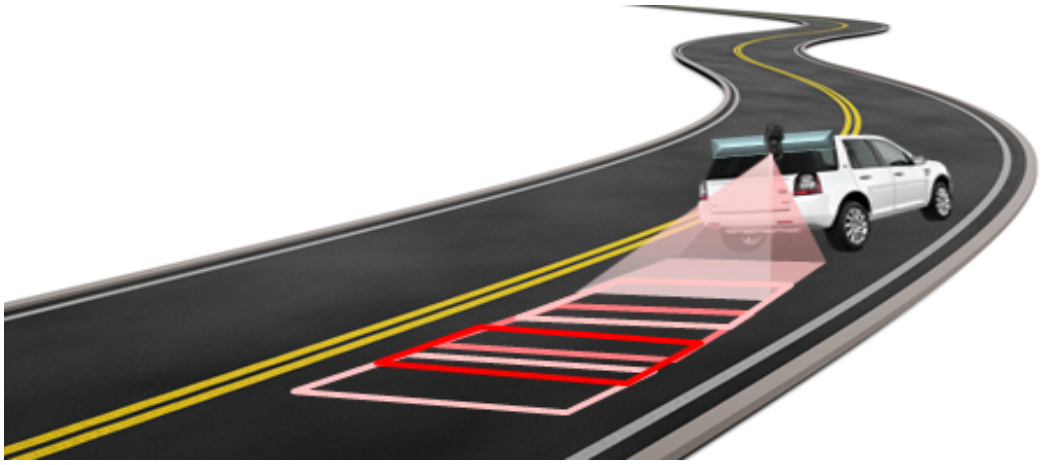


Figure 2.3: Image-based road surface mapping system with one camera

The image-based road surface mapping technique can be classified based on the number

of camera used. One of the image processing technique of getting 3D road surface model is stereo-vision. Kevin and Omar [48] designed an image-based road surface mapping system based on stereo-vision system. The road surface map generated in image-based method is performed by matching images with Normalized Cross-Correlation. It calculates a correlation coefficient of sub-arrays from two images that indicates the positive or negative correlation or non-match of two images. It also generate the 3D model surface by calculating the height of image. Marcin [45] used advanced image processing algorithm of stereo vision that it can generate 3D point clouds based on two cameras' geometry. It possible to create a road map with a resolution in millimeter. Rui et al. [17] applied a novel algorithm based on Markov Random Fields (MRF) to get disparity map of road surface and re-project the horizontal plane on the v-disparity map.

The other image-based method for road surface mapping is using one single camera, known as monocular vision [24]. On the one hand, images can stitched to generate 2D road surface map, and on the other hand, modern photometric technique, Structure From Motion (SFM), can create 3D visual representations from overlapping images. In the research report from Iowa State University [6], SFM is applied by using VisualSFM package [50] which includes algorithm of multiple views for 3D reconstruction.

Image-based road surface mapping is a cost-effective method of creating 2D/3D visual representations with road texture from overlapping images. Most of above literatures indicates that the frame rate is fixed to keep enough overlapping region between consecutive images so that there will no missing region in data collection. The main problem of image-based road surface mapping techniques is the restriction from light condition. For effectively capturing image without motion blur, the shutter speed will be short, which means the system needs higher light intensity for changing the brightness of images. Some system were intended to add external illumination such as strobe light, and some others choose

to rely on daylight. Also, to diminish the effect of shadow, the image data collection may limited to specific time and driving direction. The other main problem is the scale drift and accumulated error as the 3D point clouds from monocular vision are scaled and this will be accumulated in long distance.

2.1.3 Summary

Both laser-based and image-based techniques are powerful to provide detailed measurements of road surface, and with proper calibration they can be used in the road surface inspections, such as the analysis of road defects and the characterization of pavement roughness, and the road surface profiling. Both of these methods have their own advantages and disadvantages, and the summary of the main difference between them are list in Table. 2.1.

Table 2.1: Comparison of main differences of mobile road surface mapping techniques

Techniques	*Advantages*	*Disadvantages*
Laser-based	• High accuracy and precision in distance measurement • Not require separate illuminations	• Huge data volume and data discontinuity • Higher cost and data preprocessing is time-consuming • System will not effective when satellite is obscured
Image-based	• Continuous data collection with road surface texture • Lower cost and safer in operation	• Performance affected by illumination condition and shadows • Point clouds error depends on algorithm and has higher variance • Error is accumulated over travel distance

To build the road surface map in world coordinate, the global position of each data is the most important. From reviewed techniques, laser-based methods can generate 3D laser

scanning point cloud for 3D models of road surface, but their accuracy high rely on the position system, which is affected if satellites are blocked. Compare to this method, image-based method has potentials to reduce the rely on position system as there are overlapping between neighboring data, and it helps when positioning the road surface image. In the following section, vision-based localization techniques are reviewed.

2.2 Vision-Based Localization

From the above overview of road surface mapping system, it is obvious that image-based technique is the trend towards the mobile road surface mapping system, since the demand increasing for cost-effective in the system design and data processing. To efficiently cover wide of road surface and generate the map, this thesis chooses to focus on image-based technique on road surface mapping with one single camera. There are two key components in image-based road surface mapping: digital imaging and accurate position. This section provides a review of existed researches on vision-based localization, which employs images in localization. The past works can be classified into pure vision-based method for localization, Sec. 2.2.1, and vision and sensor fusion for localization, Sec. 2.2.2.

2.2.1 Pure Vision-Based Method of Localization

In certain region, especially in short distance, applying pure vision-based methods can give accurate location of images or other interests with geometry transformation of cameras. In the pure vision-based method of localization, images can provide abundant details of road surface or surroundings and their application in localization has been investigated intensively in the last few decades.

The basis of the pure vision-based method is to estimate the transformation between cameras, which is known as visual odometry that introduced by Nistér, Naroditsky, Bergen [37]. Visual odometry has been used for ground vehicle localization in outdoor environment, which is the similar background of mobile road surface mapping system. Visual odometry can obtain motion estimation by matching features from a sequence of images. For vision-only deployment, scale drift is a noted issue for accurate localization, especially for monocular visual odometry, as errors are accumulated overtime in long distance travel [28], like wheel odometry.

To decrease the drift error accumulated from visual odometry, Simultaneous Localization and Mapping (SLAM) is the most common fundamental applied in many techniques [11, 30, 46]. Tardif et al. [46] proposed Monocular SLAM (Mono-SLAM) which localizes the moving vehicle with 2.47% error over 2.5 km by decoupling the rotation and using epipolar constraints. Bundle Adjustment (BA) has also been used to resolve the drift error problem [18, 30], which is effective by tracking over multiple frames and highly utilized in 3D reconstruction. Frost et al. [18] incorporated the size of the objects into BA to estimate the scale, and Kurt and Motilal [30] proposed FrameSLAM which using both BA and SLAM to localize with visual imagery in large-scale.

Most of the above aforementioned methods are focus more on forward-facing cameras or has few angles to horizon. For road surface mapping, the camera is ground-facing for having more and clearer view of road surface texture information, and it has less data processing, such as road extraction and orthogonal rectification. However, the main challenge in road surface mapping with above method is that these methods are not robustly applied in low texture environment which has few significant features, like road surface, white wall and etc. Same as some indoor environment, Choi et al. [11] proposed a ceiling-view (CV) SLAM which has the similar geometry as a ground-facing camera system. They used distributed

line on the ceiling to aid the localization and reduce the computation. For outdoor road surface, as limited salient features tracked, Yanf et al. [52] proposed a monocular plane SLAM (Pop-up SLAM) which can have decent state estimation and mapping in low-texture environment. Their test has been performed in the indoor environment. On the 60 m long distance with loop, their method can has error of 0.67%.

2.2.2 Vision Sensor Fusion for Localization

The researches in pure vision-based method of localization achieved good results in both short and long distance. They improves the ability of localization, especially when GPS and other external position sensors are unavailable and unstable. However, for building a reliable map in world coordinate, global position sensor is necessary. Therefore, besides pure vision-based methods, the other common method is to integrate vision with multiple sensors and other reliable sources like map databases by using filtering or Bayesian approach. This section will review techniques that integrated vision with different sensors.

Modified Kalman filter (KF) is the most widely used technique in the integration of vision and sensors. The sensors that integrated with visual odometry for correcting the drift error are inertial measurement unit (IMU) [4, 29], Real-Time Kinematic GPS (RTK-GPS) [41, 49], wheel odometry [4], etc. One of the most popular sensor is IMU, which provides 3D measurements of both specific force and angular rate using the combination of accelerometers, gyroscopes and magnetometers. Kneip et al. proposed that the employ of vision and IMU could address the scale problem and recover relative camera motion [29]. They proposed an new Structure from Motion (SFM) approach with incorporating inertia data, which aid the computation of geometric pose.

The other common sensor that integrated with vision is Global Navigation Satellite

Systems (GNSS). It provides absolute measurement of position and time information to a passive receiver with a unique code [33]. The receiver will receive the signal's time of flight from each measured satellite and the distance to each satellite can be calculated for finding the 3D position on or near the earth surface by triangulation [43]. Nowadays, the most common GNSS, Global Positioning System (GPS) owned by the United States government, consists of up to 32 Medium Earth Orbit (MEO) satellites in 6 different orbit panels. It has been employed in both military and commercial.

There are many factors that can affect GPS accuracy, such as ionosphere and troposphere delays, signal multipath, number of satellites and etc. The error of current commercial GPS can be reduced to 3-5 m. To achieve higher accuracy, Shi et al. [44] used a rigorous sensor model in a panoramic camera for mapping the environment and a GPS for global optimization. Chen et al. [10] used monocular vision-based technique with a low-cost position sensor. Wei et al. [49] proposed the integration of vision and RTK-GPS with EKF to retrajectory the global position and proved that this method can fit ground truth better than vision-only methods. Agrawal and Konolige used commercial GPS receiver to provide absolute location information and use landmarks to improve accuracy from GPS [4]. To achieve better local position accuracy, some people attempts other sources except position sensons. Parra et al. used the street views for accurate local localization based on the relative view to the surrounding environment [40].

2.2.3 Summary

For the pure vision-based method, researchers are focus on improving the algorithm of image processing and optimizing location by reducing the match error. The performance of pure vision-based method has achieved good results in short distance and has challenged in long

Table 2.2: Comparison of main differences of vision-based localization method

Techniques	*Advantages*	*Disadvantages*
Pure vision-based	• Cost-effective and more robust in shorter distance • Without relying on motion model	• Error accumulated in a long distance • Require common features tracked over multiple frames
Vision sensor fusion	• Position corrected when visual odometry error accumulated • Less rely on initial estimation and state estimation	• More expensive with adding sensors • Performance affected by sensor quality

distance. The other class which combined vision information with other sensors to optimize position cover the weakness of vision-only method, but it also needs to solve the uncertainty from other sensors. The main advantages and disadvantages are presented in Table. 2.2.

2.3 Summary

This chapter provided an overview of the road surface mapping techniques that starts from road surface mapping system and narrows down to monocular vision-based localization of road surface images. As the cost-effective technique is required, the need of image-based road surface mapping approach is increasing. Along with this, vision-based localization techniques play a significant role in generating road surface map in global. With having pure vision-based technique that improves the estimation of motion, the integration of position sensors help to solve the scale drift issue for monocular visual odometry and reduce the dependence of state estimation. At present, there are many studies researched on vision-based localization, but due to different sensor geometry for road surface mapping and different system design, the algorithm of localizing road surface images need to be developed separately.

Chapter 3

Road Surface Mapping

Road surface mapping is an important technique to inspect the road surface condition, which can directly affect the driving safety and comfort. Vision-based road surface mapping system was chosen in the research work, as it can collect rich road surface information as described in the literature review. The proposed road surface mapping system in this research is a vehicle equipped with position sensors, speed sensors and the camera system, as shown in Fig. 3.1. The measurement of system, $\boldsymbol{z}_k^c$, includes global position measurement, $\boldsymbol{z}_k^p$, vehicle speed

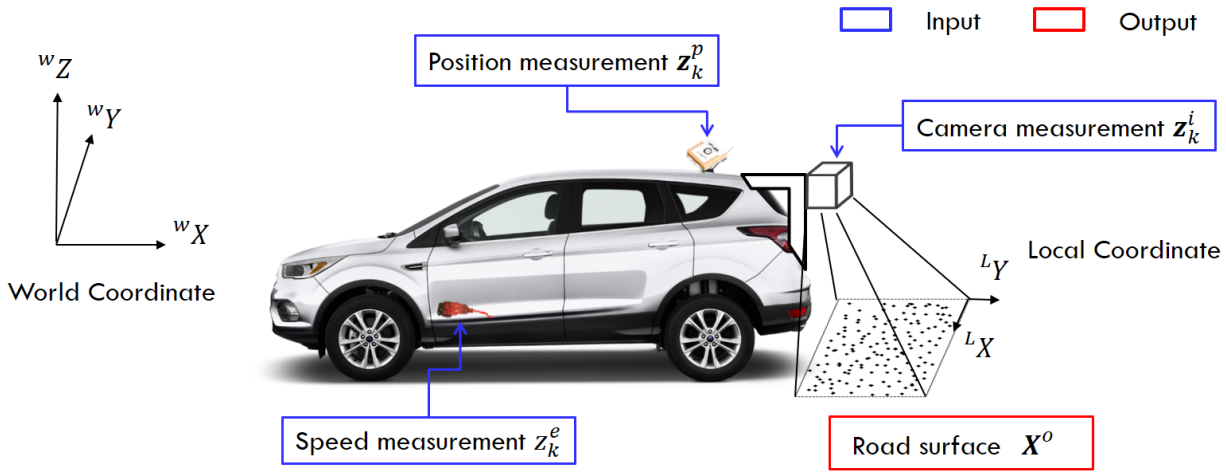


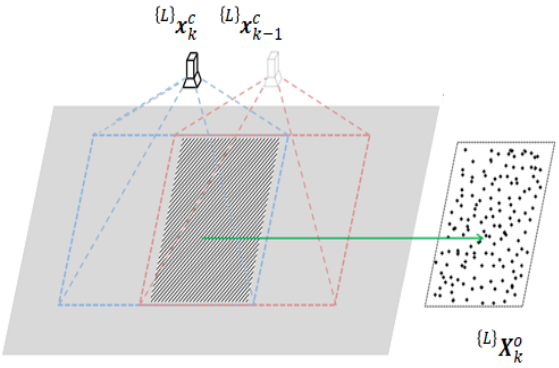
Figure 3.1: Global road surface mapping system

measurement, z_k^e, and camera measurement, $\boldsymbol{z}_k^i$. With the measurement collected along the entire route, $\boldsymbol{z}_{1:k}^c = \{\boldsymbol{z}_{1:k}^p, z_{1:k}^e, \boldsymbol{z}_{1:k}^i\}$, the designed road surface mapping system can generate the global road surface map, $\boldsymbol{X}^o$ in the world coordinate.

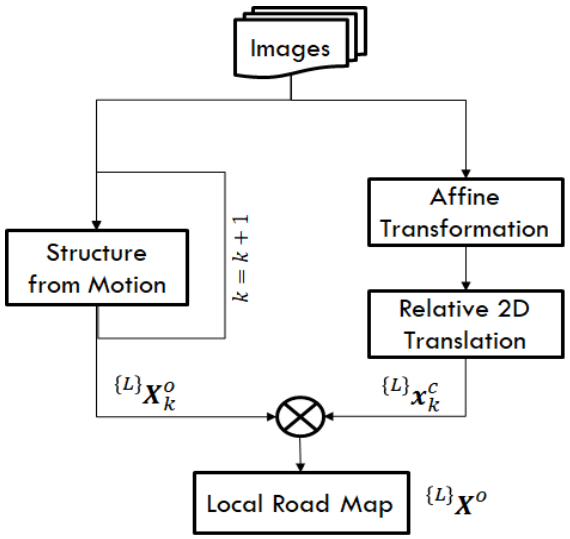
3.1 Vision-Based Road Surface Mapping in Local

The past work has demonstrated the local road surface mapping technique in image coordinate which implemented by Structure from Motion (SFM). And the local road surface map ${}^{\{L\}}\boldsymbol{X}^o$ is created by stitching road surface segments ${}^{\{L\}}\boldsymbol{X}^o_k$ along the local image plane position ${}^{\{L\}}\boldsymbol{x}^r_k$, as shown in Fig. 3.2b.

$$ {}^{\{L\}}\boldsymbol{X}^o = \{ {}^{\{L\}}\boldsymbol{X}^o_k + {}^{\{L\}}\boldsymbol{x}^r_k \}_{\forall k} \tag{3.1} $$



(a) Generate road surface point cloud with SFM



(b) Local road surface mapping flowchart

Figure 3.2: Road surface mapping in image coordinate

For each road segment at time step k, the point cloud is generated by SFM with two consecutive images, as shown in Fig. 3.2a.

$$ {}^{\{L\}}\boldsymbol{X}^o_k = \boldsymbol{f}(\boldsymbol{u}_{k-1}, \boldsymbol{u}_k, \boldsymbol{K}) \tag{3.2} $$

where $\boldsymbol{u}_{k-1}$ and $\boldsymbol{u}_k$ are the homogeneous 2D points of matched features in camera measurement $\boldsymbol{z}^i_{k-1}$ and $\boldsymbol{z}^i_k$ in camera pixel coordinate, and $\boldsymbol{K}$ is the camera intrinsic parameter

matrix that obtained by camera calibration. The fundamental of generating road surface point cloud is using triangulation with camera poses. To find the camera pose in the local coordinate, SFM using the epipolar geometry to constraint, as shown in Fig. 3.3. The essential matrix at time step k, $\boldsymbol{E}_k$, and the fundamental matrix at time step k, $\boldsymbol{F}_k$, describe the transformation between two consecutive frames and are given by

$$\boldsymbol{x}_{k-1}^T \boldsymbol{E}_k \boldsymbol{x}_k = 0 \tag{3.3}$$

$$\boldsymbol{u}_{k-1}^T \boldsymbol{F}_k \boldsymbol{u}_k = 0 \tag{3.4}$$

where he homogeneous 2D points in camera image coordinate is $\boldsymbol{x}_k \sim \boldsymbol{K}^{-1}\boldsymbol{u}_k$. By replacing it in Eq. 3.3,

$$\boldsymbol{u}_{k-1}^T (\boldsymbol{K}^{-T} \boldsymbol{E}_k \boldsymbol{K}^{-1}) \boldsymbol{u}_k = 0 \tag{3.5}$$

the relation between fundamental matrix and essential matrix is

$$\boldsymbol{F}_k = \boldsymbol{K}^{-T} \boldsymbol{E}_k \boldsymbol{K}^{-1} \tag{3.6}$$

As the fundamental matrix can be calculated by 8 or more matched points in $\boldsymbol{u}_{k-1}^T$ and $\boldsymbol{u}_k$, the corresponding essential matrix can be obtained as

$$\boldsymbol{E}_k = \boldsymbol{K}^{-1} \boldsymbol{F}_k \boldsymbol{K} \tag{3.7}$$

With applying the singular value decomposition (SVD) of the essential matrix, $\boldsymbol{E}_k = \boldsymbol{U}_k \boldsymbol{D}_k \boldsymbol{V}_k^T$, the relative rotation, $\boldsymbol{R}_{k-1,k}$, and translation, $\boldsymbol{t}_{k-1,k}$ can be obtained as

$$\begin{aligned} \boldsymbol{R}_{k-1,k} &= \boldsymbol{U}_k \boldsymbol{W} \boldsymbol{V}_k^T \\ \boldsymbol{R}_{k-1,k} &= \boldsymbol{U}_k \boldsymbol{W}^T \boldsymbol{U}_k^T \end{aligned} \tag{3.8}$$

where $\boldsymbol{U}_k$ and $\boldsymbol{V}_k$ are orthogonal matrix and $\boldsymbol{W} = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

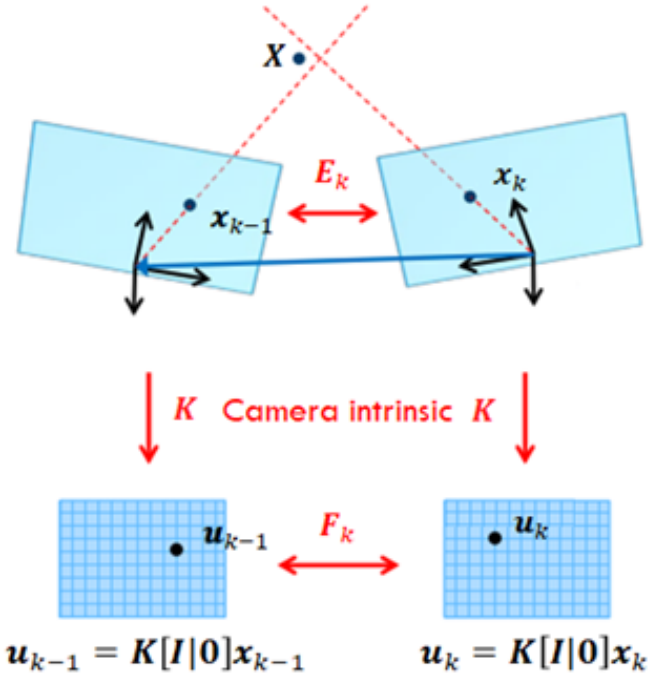


Figure 3.3: Relative camera pose estimation from epipolar geometry

In the local road surface mapping technique, for each road surface point cloud segment generation, the location of camera at the time step $k-1$ is always set at the original such that the projection at time step $k-1$ are $\boldsymbol{R}_{k-1} = \boldsymbol{I}$ and $\boldsymbol{t}_{k-1} = \boldsymbol{0}$. Then, the projection at time step k will be $\boldsymbol{R}_k = \boldsymbol{R}_{k-1,k}$ and $\boldsymbol{t}_k = \boldsymbol{t}_{k-1,k}$. The projection matrix will be represented as

$$\boldsymbol{M}_k = \boldsymbol{K}\,[\boldsymbol{R}_k|\boldsymbol{t}_k] \tag{3.9}$$

The triangulation can reconstruct the 3D road surface point by minimizing the reprojection

error so that the road surface generated at each time step is

$$ {}^{\{L\}}\boldsymbol{X}_k^o = \underset{\{{}^{\{L\}}\boldsymbol{x}_{k_j}^o\}\forall j}{\operatorname{argmin}} \sum_{i=1}^{2}\sum_{j=1}^{n} dist\left(\boldsymbol{u}_{k+i-1}^j, \boldsymbol{M}_{k+i-1}{}^{\{L\}}\boldsymbol{x}_{k_j}^o\right)^2 \tag{3.10} $$

where $\{L\}$ indicates the variable is in local coordinate, *dist* is difference function that calculate the reprojection error for each feature point. As the road surface segment at each time steps are generated at the origin, the camera positions along the image sequence is generally calculated from the translation of Affine transformation. For each feature point at time step k, the affine transformation is described as

$$ \boldsymbol{u}_k^j = \begin{bmatrix} u_k^j \\ v_k^j \end{bmatrix} = \begin{bmatrix} m_1 & m_2 \\ m_3 & m_4 \end{bmatrix} \begin{bmatrix} u_{k-1}^j \\ v_{k-1}^j \end{bmatrix} + \begin{bmatrix} t_1 \\ t_2 \end{bmatrix} \tag{3.11} $$

where $\boldsymbol{t}_{k-1,k}^{2D} = [t_1, t_2]^T$ is the translation found from Affine transformation and it can be solved with matched feature points. As the rotation matrix contains image skew and other properties, it was not used as camera transformation. The final road surface plane positions are determined by using the first image plane as the 2D reference plane

$$ {}^{\{L\}}\boldsymbol{x}_k^r = {}^{\{L\}}\ \boldsymbol{x}_{k-1}^r + \boldsymbol{t}_{k-1,k}^{2D} \tag{3.12} $$

As most local road surface mapping technique is generated to analysis the road surface macro-texture, short road surface stitched in the 2D reference plane can help to implement various road surface analysis. However, this will also limit other larger scale road surface information, such as road elevation, road pose and etc.

3.2 Camera Global Localization

For obtaining more road surface information mentioned in the previous section, the global road surface mapping is necessary which generate road surface map with geolocation information. As the camera is parallel to the ground surface with fixed height, the global road surface mapping can be simplified as camera global localization. The camera pose in world coordinate at time step k is

$$\boldsymbol{x}_k^c = [\boldsymbol{p}_k^c, \boldsymbol{r}_k^c]^T = [x_k^c, y_k^c, z_k^c, \alpha_k^c, \beta_k^c, \gamma_k^c]^T \tag{3.13}$$

where $\boldsymbol{p}_k^c = [x_k^c, y_k^c, z_k^c]$ is the camera position in world coordinate, in which x_k^c and y_k^c are the camera position in east and north direction and z_k^c is in vertical direction. $\boldsymbol{r}_k^c = [\alpha_k^c, \beta_k^c, \gamma_k^c]$ is the orientation (Euler angle) with respect to the world coordinate.

The camera global localization system can be formulated as the state equation, Eq. 3.14, and observation equation, Eq. 3.15. Assuming there is random noise for all time steps and the initial state $\boldsymbol{x}_0^c$, the image pose can be determined in the prediction and correction stages.

$$\boldsymbol{x}_k^c = \boldsymbol{f}_k(\boldsymbol{x}_{k-1}^c) + \boldsymbol{v}_{k-1} \tag{3.14}$$

where $\boldsymbol{f}_k(\cdot)$ is the state transition function and $\boldsymbol{v}_{k-1}$ is the process noise with zero mean and covariance $\boldsymbol{Q}_{k-1}$. The observation model of 3D road surface image localization is defined with observation function $\boldsymbol{h}_k(\cdot)$ which relates the current state and the observation $\boldsymbol{z}_k^c$:

$$\boldsymbol{z}_k^c = \boldsymbol{h}_k(\boldsymbol{x}_k^c) + \boldsymbol{w}_k \tag{3.15}$$

where $\boldsymbol{w}_k$ is the observation noise with zero mean and noise covariance $\boldsymbol{R}_k$.

The conventional method to solve the localization problem is using recursive Bayesian method. Kalman filter (KF) is one of this method and able to achieve the optimal state estimation. However, it is not suitable for nonlinear systems. Thus, the extended KF (EKF) is commonly used to overcome this problem by Taylor series expansion. The EKF can be a key framework of camera localization, which gives an optimal estimation of the state from system models with noises and periodically updates the state from measurements. The conventional EKF framework consists of three basic stages: initialization, prediction and correction, as shown in Fig. 3.4.

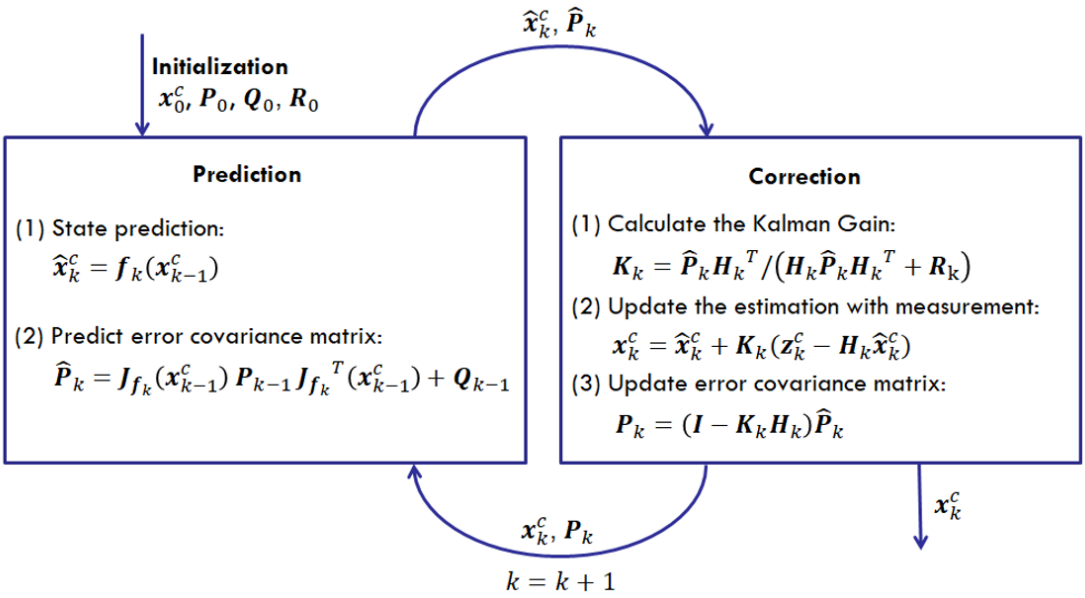


Figure 3.4: Conventional EKF framework

Prediction

In the prediction stage, nonlinear state transition function, $\boldsymbol{f}_k(\cdot)$, is used to estimate the current state of concern at time step k, $\widehat{\boldsymbol{x}}_k^c$, from its previous state, $\boldsymbol{x}_{k-1}^c$.

$$\widehat{\boldsymbol{x}}_k^c = \boldsymbol{f}_k(\boldsymbol{x}_{k-1}^c) \tag{3.16}$$

$$\widehat{\boldsymbol{P}}_k = \boldsymbol{J}_{\boldsymbol{f}_k}(\boldsymbol{x}^c_{k-1})\boldsymbol{P}_{k-1}{\boldsymbol{J}_{\boldsymbol{f}_k}}^T(\boldsymbol{x}^c_{k-1}) + \boldsymbol{Q}_{k-1} \tag{3.17}$$

where $\widehat{\boldsymbol{x}}^c_k$ and $\widehat{\boldsymbol{P}}_k$ denote the priori state vector and error covariance matrix, $\boldsymbol{x}^c_k$ and $\boldsymbol{P}_k$ are the posteriori state vector and error covariance matrix, $\boldsymbol{J}_{\boldsymbol{f}_k}$ is the Jacobian matrix, and $\boldsymbol{Q}_{k-1}$ is the covariance matrix of process noise.

Correction

In the correction stage, with having new observation or measurement $\boldsymbol{z}^c_k$, the image pose and error covariance of current state is updated:

$$\boldsymbol{x}^c_k = \widehat{\boldsymbol{x}}^c_k + \boldsymbol{K}_k(\boldsymbol{z}^c_k - \boldsymbol{H}_k\widehat{\boldsymbol{x}}^c_k) \tag{3.18}$$

$$\boldsymbol{P}_k = (\boldsymbol{I} - \boldsymbol{K}_k\boldsymbol{H}_k)\widehat{\boldsymbol{P}}_k \tag{3.19}$$

where $\boldsymbol{H}_k$ is the Jacobian matrix of the observation function, and $\boldsymbol{K}_k$ is the Kalman gain,

$$\boldsymbol{K}_k = \widehat{\boldsymbol{P}}_k{\boldsymbol{H}_k}^T(\boldsymbol{H}_k\widehat{\boldsymbol{P}}_k{\boldsymbol{H}_k}^T + \boldsymbol{R}_k)^{-1} \tag{3.20}$$

where $\boldsymbol{R}_k$ is the observation noise matrix

This conventional framework can give much more accurate location information in global than local road surface mapping, but there will be no road surface information, such as color and texture. Also, the conventional EKF model assumes that the process noise and observation noise are zero-mean Gaussian noise with fixed covariance matrix, thus it will work well for tuned model. However, if the error are not white noises, the filter will not perform as expected when observation environment change. Also, as the camera is ground-facing and it limited view, the frame rate will be much higher than observation rate, especially when vehicle is on highway.

3.3 Proposed 3D Road Surface Mapping with Geolocation

As the limitation of conventional method in local road surface mapping and global camera localization, the 3D global road surface mapping technique is proposed which generates road surface map with geolocation. The main framework of the proposed system is shown in Fig. 3.5, which combing the techniques of local road surface mapping and camera global localization.

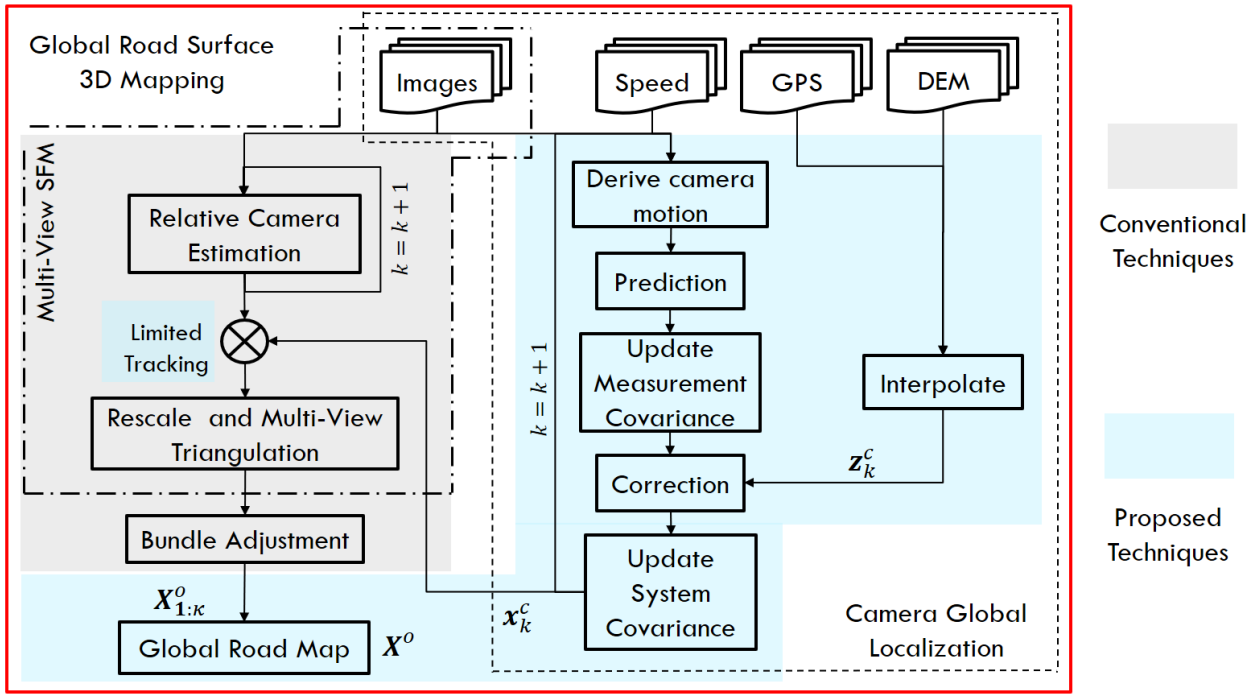


Figure 3.5: Proposed 3D road surface mapping framework

The conventional techniques used in the proposed method is indicated with gray shadow and the proposed new techniques are indicated with blue shadow. To accurately localize camera in global, the designed system applied a modified EKF to integrate the vision from a ground-facing camera and a position sensor with a digital map. To solve

the uncertainty of the observation, the modified version of conventional EKF need to be included. One attempt to detect the filter divergence upon EKF is to add a fuzzy logic controller to detect the bias for preventing divergence [42]. The modified version of EKF is proposed which autonomously tunning process noise covariance to improve performance [12]. However, even though adaptively adjusting the noise covariances to avoid divergence, the conventional prediction model assumes smooth camera motion and constant velocities, which means unexpected error can occur during sudden changes in motion [5]. In the proposed method, the adaptive EKF (AEKF) was used to reduce the error from dynamic environment by incorporating the prior knowledge from vision and new observations from the integration of GPS and digital map to localize camera in world coordinate. Also, comparing low rate measurement with required high rate state update for road surface mapping application, the correction stage is not efficient. The interpolate method is introduced to interpolated more measurements for sufficient correction in Chapter 4.

The new road surface mapping system is proposed by using multiple views for triangulation. Although this technique has been applied in many other field, in the road surface mapping, the object, road, belongs to low-texture environment which has the significant error in the vertical direction respect to the ground surface. To reduce the error vertical to road surface the proposed technique limits the tracked feature and reduce the reprojection error with Bundle Adjustment (BA) in each short road segment. The final 3D road surface map with geolocation is generated with combining the optimized local road surface map and camera global localization results.

Chapter 4

AEKF-Based Camera Global Localization

Some powerful techniques can achieve vision-based localization with front-view camera, but they depend on the accurate estimation of frame-to-frame motion and faces a drawback in low-texture environment, such as road surface. Since the density of features in low-textured environment and the field of view (FOV) for ground-facing camera are limited, in this chapter, we propose an AEKF-based model for camera global localization in road surface mapping system for optimizing camera 3D pose in world coordinate. The proposed technique is built up based on the conventional EKF model with modifications in prediction, correction and adaptive covariance matrix. This chapter will demonstrate the details for each part in sections.

4.1 Adaptive Noise Covariance Matrix

The environment keeps changing as driving on the road. If the system is in the stable environment, the conventional properly modeled EKF will perform as desired. The covariance matrix in conventional EKF is always a constant matrix based on the sensor accuracy. However, for global position sensor, the GPS accuracy and precision are vary in different environment so that the uncertainty from GPS measurement should be different. Also, the

motion estimation from visual odometry accuracy may change and be affected by different light conditions. Therefore, how to properly calculate the covariance matrices is important as they have highly affects on the performance [36].

There are several types of adaptive method which include scaling the covariance, multi-model adaptive estimation and adaptive stochastic modeling. In this thesis, adaptive estimation of the covariance matrices as vehicle moving is investigated. The adaptive method includes innovation-based and residual-based, where residual is the difference between observe value and the estimated value and innovation is the difference between the true value and the estimated value.

4.1.1 Adaptive Estimation of Measurement Noise Matrix

Although the accuracy of GPS is stable in open space, it is vary when the system is operated in different environment. In order to get higher position accuracy, the measurement noise covariance is adaptively adjusted based on residual, $\boldsymbol{e}_k$, which perform more accurately for low-cost sensors [23].

$$\boldsymbol{e}_k = \boldsymbol{z}_k^c - \boldsymbol{H}\widehat{\boldsymbol{x}}_k^c \tag{4.1}$$

The measurement noise covariance matrix is updated based on the covariance of residual to reduce the bias and estimation error. The above equation can be converted by finding the variance for each element in equation:

$$E\left[\boldsymbol{e}_k\boldsymbol{e}_k^T\right] = E\left[\boldsymbol{z}_k^c\boldsymbol{z}_k^{cT}\right] - \boldsymbol{H}E\left[\widehat{\boldsymbol{x}}_k^c\widehat{\boldsymbol{x}}_k^{cT}\right]\boldsymbol{H}^T \tag{4.2}$$

where $E[\cdot]$ is the symbol represents the covariance matrix which computed the average of $\boldsymbol{e}_k\boldsymbol{e}_k^T$. To obtain the covariance matrix of the residual, the above equation can be derived as

following based on the residuals in the last m time steps, $\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{e}_{k-i}\boldsymbol{e}_{k-i}^{T}$. And the covariance of $\boldsymbol{z}_k$ and $\widehat{\boldsymbol{x}}_k$ in the above equation can be replaced by $\boldsymbol{R}_k$ and $\widehat{\boldsymbol{P}}_k$.

$$\boldsymbol{R}_k = \frac{1}{m}\sum_{i=1}^{m}\boldsymbol{e}_{k-i}\boldsymbol{e}_{k-i}^{T} + \boldsymbol{H}_k\widehat{\boldsymbol{P}}_k{\boldsymbol{H}_k}^{T} \tag{4.3}$$

The measurement noise covariance will be updated based on the recent environment and calculated before each correction steps.

4.1.2 Adaptive Estimation of Process Noise Matrix

The prediction from visual odometry is stable; however, the light condition is changed as vehicle moving in different environment and affects the results. To overcome the prediction error from visual odometry, the process noise covariance is also adaptively adjusted by innovation, the difference between the actual value and predicted value.

$$\boldsymbol{d}_k = \boldsymbol{x}_k^c - \widehat{\boldsymbol{x}}_k^c \tag{4.4}$$

With applying the same strategy used for measurement noise covariance, the process covariance matrix updated based on innovations in the last m time steps is

$$\boldsymbol{Q}_k = \frac{1}{m}\sum_{i=1}^{m}\boldsymbol{d}_{k-i}\boldsymbol{d}_{k-i}^{T} + \boldsymbol{P}_k - \boldsymbol{J}_{\boldsymbol{f}_k}(\boldsymbol{x}_{k-1}^c)\boldsymbol{P}_{k-1}{\boldsymbol{J}_{\boldsymbol{f}_k}}^{T}(\boldsymbol{x}_{k-1}^c) \tag{4.5}$$

4.2 Camera Pose Prediction with Image Shift

A traditional EKF model uses synchronized data to optimize position, so the most common method is using motion model in prediction stage. However, in a motion model, which

uses constant velocities and assumes smooth camera motion, error can occur during sudden changes in motion. Alcantarilla et al. proposed to use visual odometry as prior in a recursive filter and proved that it improve the robustness of localization [5]. The visual odometry from a ground-facing camera system is designed to be estimated by image shift between consecutive frames [7, 38] to reduce the effect from low texture environment. Nourani-Vatani and Borges used a simplified motion model with image shift to improve accuracy by multi-template correlation, and achieved a translation error of 5% over 5 km [38]. Therefore, unlike the conventional EKF, the proposed AEKF-based camera global localization uses the proposed visual odometry as priors to EKF for prediction and corrects states by interpolated GPS readings with using adaptive covariance matrix.

Instead of using standard motion model with constant velocity, visual odometry is applied as priors. Same as conventional EKF, $\widehat{\boldsymbol{x}}_k^c$ and $\widehat{\boldsymbol{P}}_k$ are the priori state vector and error covariance matrix, $\boldsymbol{x}_k^c$ and $\boldsymbol{P}_k$ are the posteriori state vector and error covariance matrix. By taking advantaging of the ground-facing camera geometry and the fixed camera height with respect to the ground, the camera pose can be estimated by image shift to overcome the vertical noise in long distance. Therefore, $\boldsymbol{f}_k$ is a function of visual odometry priors,

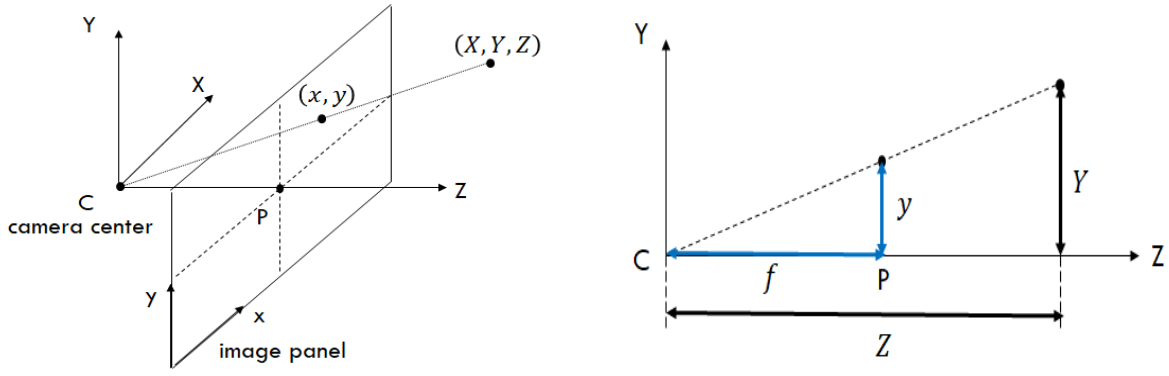
$$\widehat{\boldsymbol{x}}_k^c = \boldsymbol{f}_k(\boldsymbol{x}_{k-1}^c) = \begin{bmatrix} \widehat{\boldsymbol{p}}_k^c \\ \widehat{\boldsymbol{r}}_k^c \end{bmatrix} = \begin{bmatrix} \boldsymbol{p}_{k-1}^c + rotm(\boldsymbol{r}_{k-1}^c)\widetilde{\boldsymbol{t}}_{k-1,k}^c \\ eul(\widetilde{\boldsymbol{R}}_{k-1,k}^c rotm(\boldsymbol{r}_{k-1}^c)) \end{bmatrix} \tag{4.6}$$

where $\widetilde{\boldsymbol{R}}_{k-1,k} \in \mathbb{R}^{3x3}$ and $\widetilde{\boldsymbol{t}}_{k-1,k} \in \mathbb{R}^3$ are the relative image transformation from visual odometry. And the Euler angles are converted to rotation matrix with function $rotm(\cdot)$ before calculating and converted back to Euler angle using $eul(\cdot)$ function. $\boldsymbol{J}_{\boldsymbol{f}_k}$ is the Jacobian matrix and $\boldsymbol{H}_k$ is an identity matrix here since the proposed method interpolates both position and rotation based on 3D B-spline curve of GPS readings, which will be shown in Sec. 4.3.

The visual odometry from a ground-facing camera system is designed to be estimated by image shift between consecutive frames. This section provides the fundamentals and details of image processing in prior for image shift estimation. Initial image shift estimation is based on the rigid image transformation which is applied with ARANSAC to remove the outliers. Also, the addition motion constraint from frame rate control is used to avoid the static estimation from vehicle moving shadow. Consequently, the estimate movement will be converted to world coordinate.

4.2.1 Camera Calibration

Camera calibration can estimate camera intrinsic, extrinsic and lens distortion parameter which can not only help to remove the effects of lens distortion, but also improve the estimation of the camera motion, or 3D reconstruction from images. As the camera system is fixed on vehicle which has the constant distance between the camera and the ground surface. Therefore, the extrinsic of calibrated camera can be used to find the camera height. The standard camera model for calibration is commonly assumed to be a pinhole camera model as shown in Fig. 4.1a.



(a) Pinhole camera model (b) The side view of pinhole camera model

Figure 4.1: Camera model

Camera center, C, is the optical center, and P is the principal point on the image panel. The position of interest point in world coordinate is represented as (X, Y, Z) and its corresponding position on the mirrored image panel is (x, y). The relation between the position in the world coordinate and image coordinate is shown in Eq. 4.7 with homogeneous coordinate.

$$\lambda \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \mathbf{K} \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0} & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \tag{4.7}$$

where λ is the arbitrary homogeneous coordinates scale factor, $\mathbf{K}$ is the camera intrinsic matrix defined in Eq. 4.8, and $\mathbf{R}$, $\mathbf{t}$ consist of camera extrinsic matrix, which can be used to calculate the camera height respects to the ground. Camera intrinsic matrix describes the focal length, f, principal point offset, $c = (c_x, c_y)$, and skew parameter, s. Focal length is the distance between the optical center and image panel and vary in two directions as f_x and f_y. Fig. 4.1b illustrates the focal length in y-direction.

$$\mathbf{K} = \begin{bmatrix} f_x & s & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{4.8}$$

Besides intrinsic matrix, which does not contain lens distortion with assuming ideal pinhole camera model, distortion parameters should be considered in camera calibration. Distortion includes radial distortion and tangential distortion. Radial distortion occurs when light rays bend and tangential distortion occurs when the lens and image panel are not parallel. The radial distortion includes barrel distortion and pincushion distortion which is

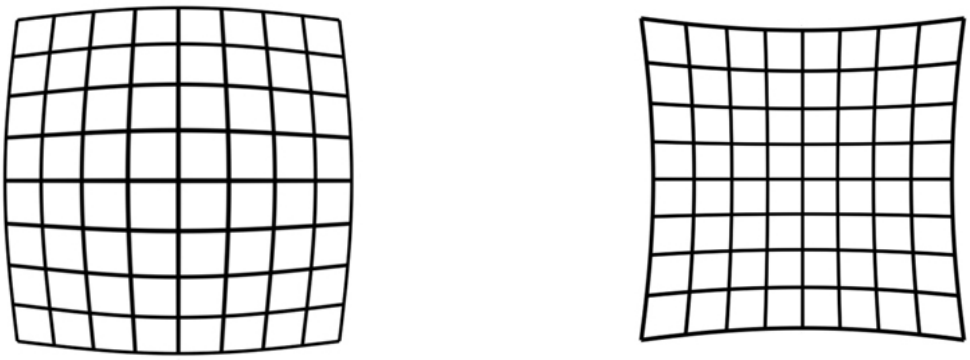
shown in Fig. 4.2. Radial distortion parameters are

$$\begin{aligned} d_{rx} &= x(1 + k_1 * r^2 + k_2 * r^4 + k_3 * r^6) \\ d_{ry} &= y(1 + k_1 * r^2 + k_2 * r^4 + k_3 * r^6) \end{aligned} \tag{4.9}$$

and tangential distortion parameters are

$$\begin{aligned} d_{tx} &= x + [2 * p_1 * x * y + p_2 * (r^2 + 2 * x^2)] \\ d_{ty} &= y + [2 * p_2 * x * y + p_1 * (r^2 + 2 * y^2)] \end{aligned} \tag{4.10}$$

where d_{rx} and d_{ry} are radial distortion parameters along x and y directions, d_{tx} and d_{ty} are tangential distortion parameters along x and y directions, (x, y) is the undistorted pixel location, $r^2 = x^2 + y^2$ and k_1, k_2, k_3, p_1 and p_2 are distortion coefficients of the lens. The calibration results of above mentioned parameters will be shown in Sec.6.3.



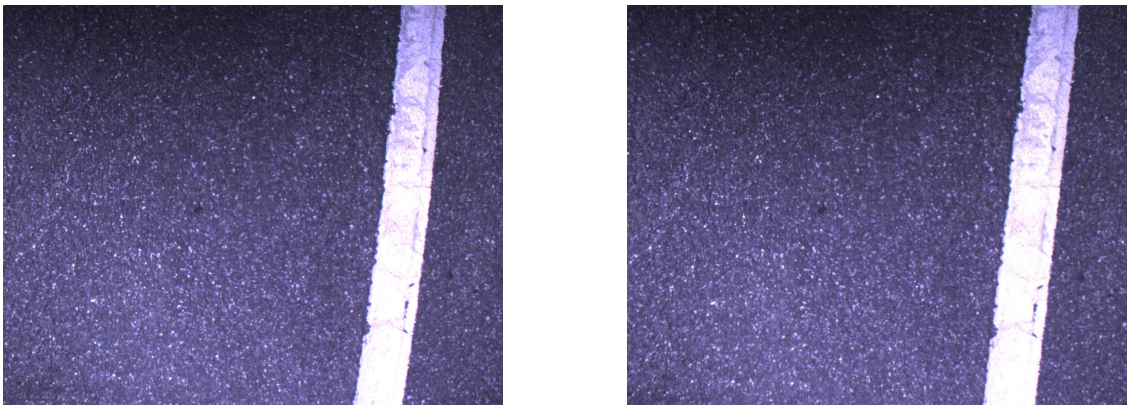
(a) Barrel distortion

(b) Pincushion distortion

Figure 4.2: Radial Distortion Type

One example of road surface image calibration is shown in Fig. 4.3. It's obvious that the image before calibration, Fig. 4.3a, contains distortion, especially on the edge of the image. The white road marking near the image edge is distorted so that the line is not straighted. After calibration, the corresponding line in the calibrated image, Fig. 4.3b, is

corrected.



(a) Image before calibration (b) Image after calibration

Figure 4.3: Example of road surface calibration

The calibration algorithm is applied to find the above mentioned camera parameters and estimate the camera height by extrinsic matrices. The lens distortion is assumed to be zero to initialize the intrinsic of camera and extrinsic of each calibration image, then all parameters are optimized by using Levenberg-Marquardt algorithm [22, 54]. From the chosen calibration checkerboard images, we could have a set of extrinsic parameters which represent the related distance between camera and checkerboards as shown in Fig. 4.4. The calibration checkerboard is flatted and put on the ground or with small angles to calibrate for camera parameters. The camera height, h, is calculated by averaging the distance of the checkerboards on the ground related to the camera, as shown with solid pattern in Fig. 4.4.

4.2.2 Feature Detection and Matching

A good image feature detection and matching algorithm is important for the estimation of motion. One of the traditional technique is Canny edge detector. It applied different filters to the smoothed image for detecting the edge in different directions. And it uses non-maximum suppression and threshold to find more accurate edges in the image. There are

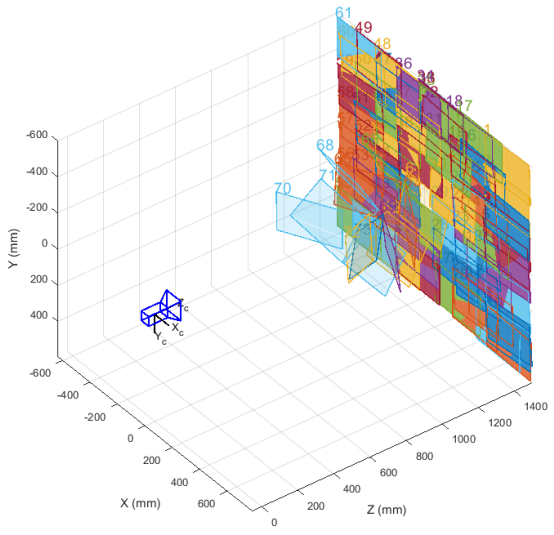
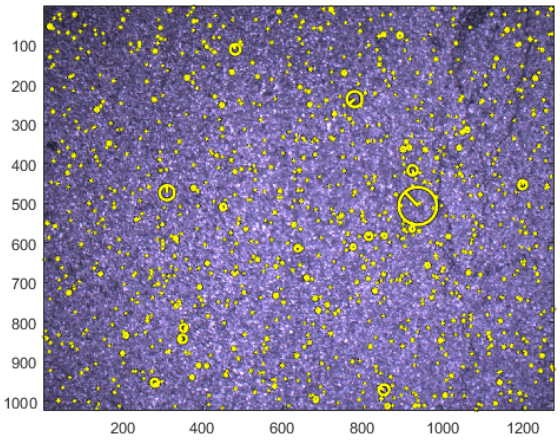


Figure 4.4: The relative positions between a camera and calibration checkerboards

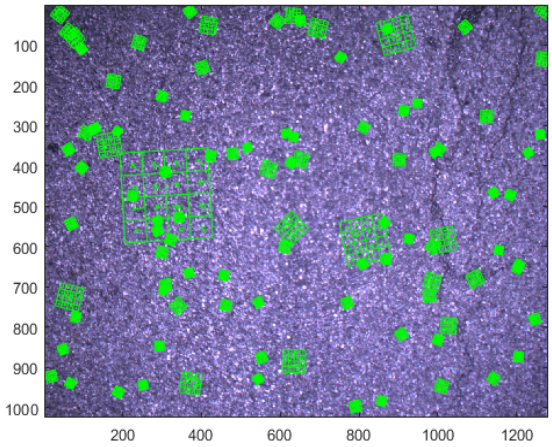
also some other techniques, such as Fast Corner Detector, Harris Corner detector and etc. These techniques has been shown the efficiency and good performance for many objects. As the interest object in this thesis does not have clear edges, the other type of method, scale-invariant feature transform (SIFT) [34], shows more robust feature detection and matching in low texture environment.

The state of the art approach of feature detector and descriptor, SIFT, is proposed by David Lowea. The detector find the keypoints by using high dimensional vectors to represent the image gradient and combining with the local region of the image. The detected keypoints are robust in extracting small objects among clutter which will be efficient in our application. In this thesis, we use the implementation of the SIFT detector provided by Vedaldi and Fulkerson [47]. As shown in Fig. 4.5, the raw result from the example road surface image has more than 10000 keypoints detected which is much more than other method. Besides the position, the detector can also detect the orientation and scale. The scale is represented based on the size of circle, and there is a line vector indicates the orientation, as shown in

Fig. 4.5a. The SIFT algorithm also provide the descriptor for all keypoints, as shown in Fig. 4.5b.



(a) Random selected keypoints with orientation and scale



(b) Random selected keypoints with descriptor

Figure 4.5: SIFT applied in a road surface image

4.2.3 Near Planer Camera Transformation

To improve the motion estimation in low texture environment, we make a simple assumption that we constrain our search for correspondences between I_k and I_{k-1}, by assuming that the relative elevation does not change significantly between neighboring images. Therefore, the movement can be assumed only in 2D. Based on the evaluated results of frame rate control in Sec. 6.3.2, the translation in vehicle heading direction between consecutive frames is around 0.27 meters. Considering the shift between consecutive image along the heading direction is small, it is reasonable to assume that the movement between consecutive frame is in a 2D plane which has small change in elevation with respect to the previous frame. The image transformation is estimated by the matched feature positions. The relative rotation $\boldsymbol{R}^c_{k-1,k}$ and translations $\boldsymbol{t}^c_{k-1,k}$ is optimized by least square rigid motion using singular value decomposition (SVD) [13].

Between two consecutive images, I_k and I_{k-1}, there are N keypoints are detected and matched. The i^{th} feature's corresponding position in two image coordinates are $\boldsymbol{p}_k^i$ and $\boldsymbol{p}_{k-1}^i$. The optimized transformation is found by minimized the sum error when applied transformation for all features.

$$\left(\boldsymbol{R}_{k-1,k}^c, \boldsymbol{t}_{k-1,k}^c\right) = \underset{\boldsymbol{R}\in\mathbb{R}^{2x2}, \boldsymbol{t}\in\mathbb{R}^2}{\operatorname{argmin}} \sum_{i=1}^{N} \left\| \boldsymbol{R}\boldsymbol{u}_{k-1}^i + \boldsymbol{t} - \boldsymbol{u}_k^i \right\|^2 \tag{4.11}$$

As there would be many combination of translation and rotation, to simplified the problem, the estimation of the optimal rotation can be found first by centralizing the matched feature point positions. The centralized i^{th} feature's position for image $\boldsymbol{I}_{k-1}$ and $\boldsymbol{I}_k$ is represented as $\boldsymbol{c}_{k-1}^i$ and $\boldsymbol{c}_k^i$,

$$\boldsymbol{c}_{k-1}^i = \boldsymbol{u}_{k-1}^i - \overline{\boldsymbol{u}}_{k-1} \qquad \boldsymbol{c}_k^i = \boldsymbol{u}_k^i - \overline{\boldsymbol{u}}_k \tag{4.12}$$

where $\overline{\boldsymbol{u}}_{k-1}$ and $\overline{\boldsymbol{u}}_{k-1}$ is the center position of all features calculated by averaging the position. Then, the optimized rotation is simplified as following.

$$\boldsymbol{R}_{k-1,k}^c = \underset{\boldsymbol{R}\in\mathbb{R}^{2x2}}{\operatorname{argmin}} \sum_{i=1}^{N} \left\| \boldsymbol{R}\boldsymbol{c}_{k-1}^i - \boldsymbol{c}_k^i \right\|^2 \tag{4.13}$$

After expanding the right side of Eq. 4.13 and removing the constant parts, the rotation equation is

$$\mathbf{R}_{k-1,k}^c = \underset{\boldsymbol{R}\in\mathbb{R}^{2x2}}{\operatorname{argmin}} \sum_{i=1}^{N} -2\boldsymbol{c}_k^{i\,T}\boldsymbol{R}\boldsymbol{c}_{k-1}^i = \underset{\boldsymbol{R}\in\mathbb{R}^{2x2}}{\operatorname{argmax}}\, trace(2\boldsymbol{R}\boldsymbol{C}_{k-1}^T\boldsymbol{C}_k) \tag{4.14}$$

where $\boldsymbol{C}_{k-1}$ and $\boldsymbol{C}_k$ are the diagonal matrix with having centered feature position on the diagonal, and *trace* is the trace of a square matrix, that summing the elements on the diagonal. Then with decomposing $\boldsymbol{C}_{k-1}^T\boldsymbol{C}_k$, the correlation matrix, by SVD, the optimal

rotation is

$$\boldsymbol{R}^{c}_{k-1,k} = \boldsymbol{V}_k \boldsymbol{\Sigma}_k \boldsymbol{U}_k^T \tag{4.15}$$

where $\boldsymbol{\Sigma}_k$ is the reflection matrix to check if there is potential reflection in addition to rotation.

$$\boldsymbol{\Sigma}_k = \begin{bmatrix} 1 & 0 \\ 0 & det(\boldsymbol{V}_k \boldsymbol{U}_k^T) \end{bmatrix} \tag{4.16}$$

Once the optimal rotation is found, applied it to the center of features to find the translation $\boldsymbol{t}_{k-1,k}$.

$$\boldsymbol{t}^{c}_{k-1,k} = \overline{\boldsymbol{u}}_k - \boldsymbol{R}^{c}_{k-1,k} \overline{\boldsymbol{u}}_{k-1} \tag{4.17}$$

Beyond this point, the initial estimation of image transformation is calculated with SVD. However, this will not promise the good estimation in any environment, especially when there are low quality features detected in low texture environment. The improvement of image transformation optimization is proposed in the following section.

4.2.4 Transformation Optimization by Adaptive RANSAC with Motion Constraint

The common image shift can guarantee the estimation based on good feature detection and matching; however, in real world, there will be outliers caused by the incorrectly matched features. Therefore, Random Sample and Consensus (RANSAC) is applied to remove the outliers for getting better results. The outliers ratio is different from frame to frame as most of the outliers' are from shadows, and shadow region varies with light direction. Thus, the proposed AEKF-based localization method adaptively adjusts the outliers ratio for each iteration in RANSAC with repeat times adjusted until convergence, which is called adaptive

RANSAC (ARANSAC) [35].

For images collected from ground-facing cameras, moving shadow is the main error that could cause wrong outliers removal, which results in a wrong motion estimation even removed noises by applying ARANSAC. To differentiate from incorrect matches caused by moving shadow, a threshold, *thres*, is applied over the inlier ratio in ARANSAC as shown in Fig. 4.6. Besides this, as the movement between two consecutive images is controlled based on the vehicle speed, the addition of motion constraint is included in ARANSAC by evaluating lower bound of image translation, t_{OBD}. Using this approach, the effects of mismatched features are negated from moving vehicle shadows. The threshold and lower bound in the proposed system are selected as 70% and 200.

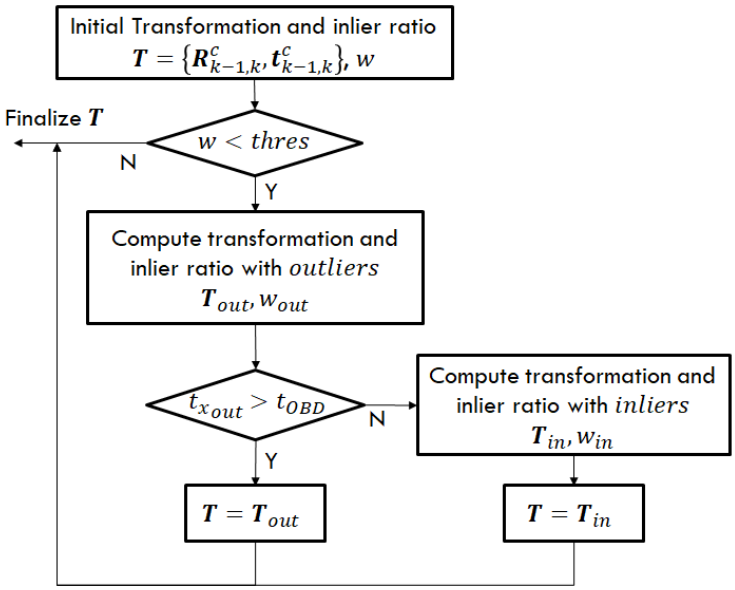
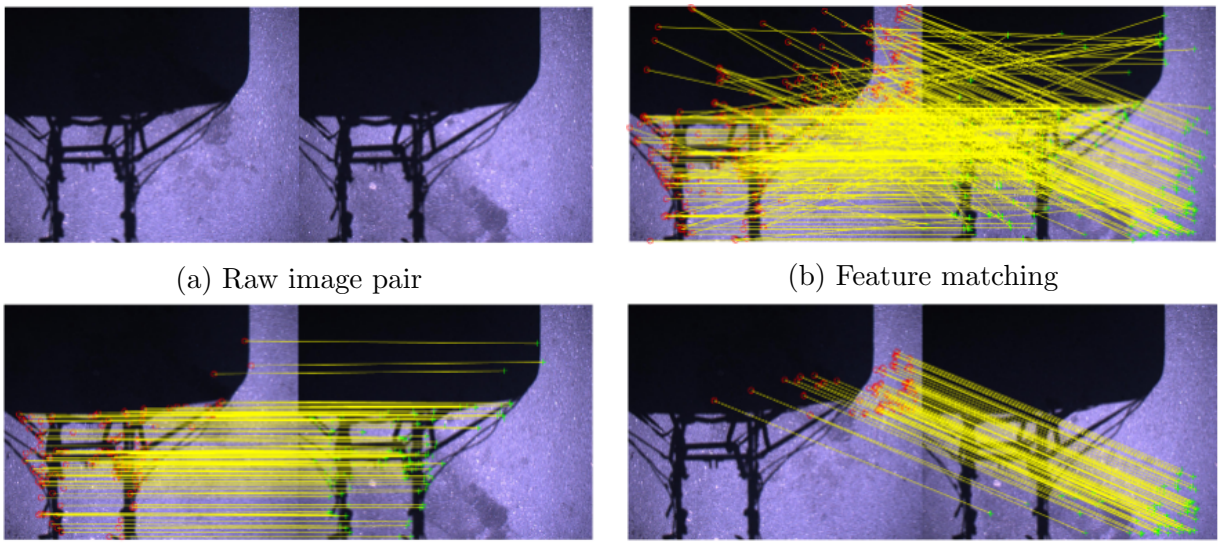


Figure 4.6: The proposed image shift optimization method

Different sources of shadow are big concerns for motion estimation using as mentioned in some previous similar works [38]. The main error of motion estimation is from moving shadow from vehicle. An example of how moving shadow affects the motion estimation is shown in Fig. 4.7, which consist of image pairs that previous image on the left and the

second image on the right. In Fig. 4.7a, the pair of image show the raw images of two consecutive images. The selected pair has a large region covered by vehicle shadow. Ignoring the shadow region, the correct movement can be estimated by the corresponding features from the spots on the road surface. Some shadows experience a lot of mismatches, as shown in Fig. 4.7b. The shadow region is too dark to find feature, so there are barely no features are detected. However, the shadow from the system is wrong detected as edge so that there are many features detected and matched, which shown as horizontal lines. Application of ARANSAC, even though significant remove the noise, sometimes it is not enough to provide accurate results when too many bad matching exist. In this example, ARANSAC results in a static vehicle motion estimate, as shown in Fig. 4.7c. From the initial feature matching in Fig. 4.7b, we find that there are two large groups of matches. One is from matching the edge of shadows, and the other one is from matching the road surface features. To differentiate these from incorrect matches, the proposed ARANSAC applies a threshold over the inlier ratio in ARANSAC to detect the potential of large wrong matches from shadow. Besides that, as the movement between two consecutive images is controlled based on the vehicle speed, the proposed algorithm also adds motion constraint in ARANSAC by evaluating translation. Using this approach, we successfully negate the effects of mismatched features in moving vehicle shadows, as shown in Fig. 4.7d. And the image transformation is updated from the corrected matches.



(a) Raw image pair (b) Feature matching

(c) ARANSAC without motion constraint (d) ARANSAC with motion constraint

Figure 4.7: Bad matching caused by moving shadow from the vehicle and optimized by the proposed ARANSAC method

4.2.5 Image Pose Prediction and Residual

Once the proposed algorithm optimizes the motion with centralized data, the corresponding transformations in 3D for the process model of proposed AEKF model are

$$\widetilde{\boldsymbol{R}}^{c}_{k-1,k} = \begin{bmatrix} \boldsymbol{R}^{c}_{k-1,k} & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix} \qquad \widetilde{\boldsymbol{t}}^{c}_{k-1,k} = r\boldsymbol{R}_{cv} \begin{bmatrix} \boldsymbol{t}^{c}_{k-1,k} \\ 0 \end{bmatrix} \tag{4.18}$$

where s is the ratio of pixel to meter calculated by calibrated fixed camera height h, field of view (FOV) and image size, and $\boldsymbol{R}_{cv}$ is the transformation from camera frame to world frame. The ratio s is defined as $l/1024$, where l is the image length along the vehicle heading direction. With having the camera height in calibration,

$$l = 2h * \tan(\frac{\alpha}{2}) \tag{4.19}$$

where α is the field of view along the vehicle heading direction as shown in Fig. 4.8

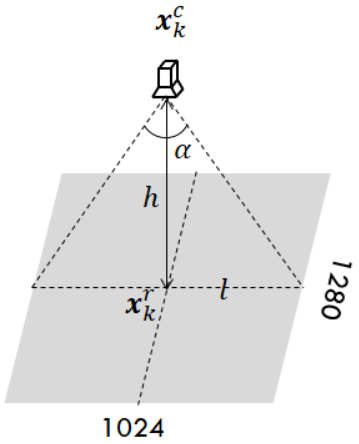


Figure 4.8: Camera pixel to meter scale

$\widetilde{\boldsymbol{R}}^c_{k-1,k}$ and $\widetilde{\boldsymbol{t}}^c_{k-1,k}$ are updated at each time step and replaced in Eq. 4.6 for the proposed process model. Unlike the traditional motion model, it can reduce the drift in sudden motion over time by keeping the overlapping region consistent between frames. This model also negate errors caused by moving vehicle shadow in visual odometry.

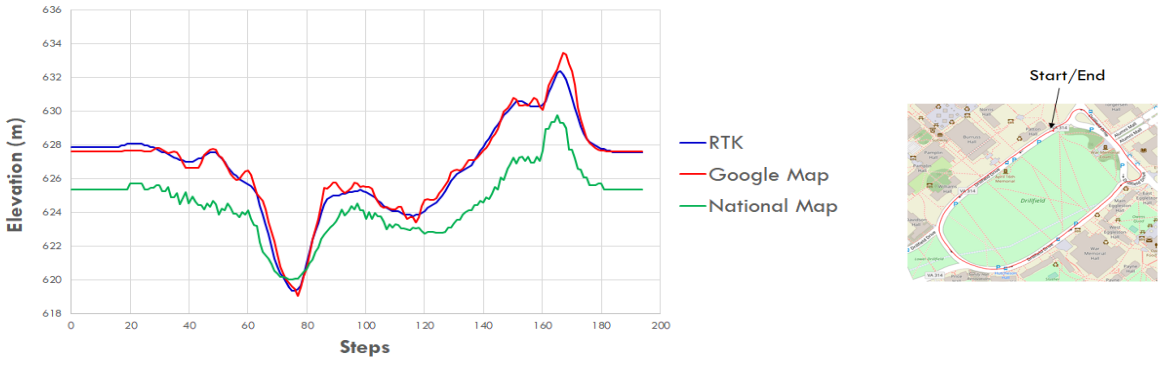
4.3 Interpolate Camera Pose for Correction

In the process model, the prediction in small distance with visual odometry is accurately estimated. Although the relative elevation change is small in short image sequences, the elevation in long distance will change as the road longitude profile is not always straight. Therefore, it is necessary to add additional position sensors to correct the position in long distance. Also, as the image capturing system is dynamically control over vehicle speed, there will be various number of images between two GPS readings and the GPS update rate is much smaller than image frame rate. In this thesis, the proposed algorithm challenges the sparse low-accuracy GPS data with combining elevation data from digital map to correct

the accumulated error from image transformation over long distance.

4.3.1 GPS 3D B-spline Curve and Measurement Interpolation

To reduce the accumulate position and rotation error in visual odometry, GPS data is used in the observation model for correcting image pose. Since the elevation data read from GPS is unreliable, the corresponding elevation from 3D road terrain map database is collected to improve the global position accuracy [14]. There are many 3D road terrain map available and shared to public, such as Google Map and National Map. Based on a field test, this thesis choose to use the elevation data read from Google Map. As shown in Fig. 4.9a, it is more accurate since it is closer to RTK GPS readings, an accurate global position sensor, along the test route, the red curve, in Fig. 4.9b. If considering the RTK GPS reading as ground truth, the mean elevation error of Google Map is 0.3 m and the standard deviation is 0.4058. In the rest content, all GPS readings indicate 3D position that the longitude and latitude read from the GPS and elevation read from digital map.



(a) Comparison of elevation from RTK GPS and Maps (b) Field test route

Figure 4.9: Field test for selecting 3D road terrain digital map

Sparse (low bandwidth 1 Hz) global position is insufficient for correcting the error in high image frame rate. By introducing a GPS 3D B-spline curve method, the corresponding

GPS data for each image can be interpolated. To generate a 3D B-spline curve, control points, GPS points $\{\boldsymbol{z}^c_{1:k}\}$ with corresponding elevation data is noted as $\boldsymbol{g} \in \boldsymbol{G}$, is selected by a square slide window with width $2 * wid$ around estimated image position $\widehat{\boldsymbol{x}}^c_k$. The size of slide window is also automatic adjusted based on the GPS data density in the window. If there is not enough GPS readings to generate curve, the slide window size will be increased until the curve generated. GPS logs set $\boldsymbol{G} = \{\boldsymbol{g}_0, \boldsymbol{g}_1, \cdots \boldsymbol{g}_n\}$ with $N = \{0, 1, \cdots n\}$ and the selected points are

$$\begin{aligned}\left\{\boldsymbol{g}_{m^i_k} \mid m^i_k \in N \wedge \left(\widehat{x}^c_k - wid \leq g_{x m^i_k} \leq \widehat{x}^c_k + wid\right)\right. \\ \left. \wedge \left(\widehat{y}^c_k - wid \leq g_{y m^i_k} \leq \widehat{y}^c_k + wid\right)\right\}\end{aligned} \tag{4.20}$$

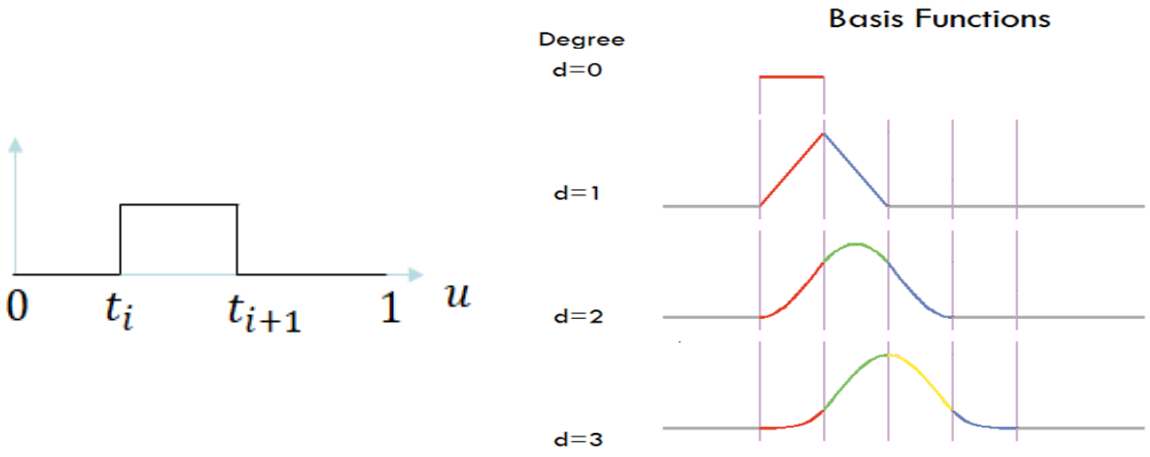
And control point set consists of (n_{cp}+1) points: $\left\{\boldsymbol{g}_{m^0_k}, \boldsymbol{g}_{m^1_k}, \cdots \boldsymbol{g}_{m^{ncp}_k}\right\}$ with uniformly-spaced knot vector $\boldsymbol{t}_{kv}$ with $(n_{kt} + 1)$ knots for d^{th} degree B-spline curve.

$$\boldsymbol{t}_{kv} = \begin{bmatrix} t_1 & t_1 & \cdots & t_{n_{kt}} \end{bmatrix} \tag{4.21}$$

The B-spline curve function is defined by a series of basis function in different degrees:

$$\begin{aligned} B_{i,0}(u) &= \begin{cases} 1, & t_i \leq u \leq t_{i+1} \\ 0, & \text{otherwise} \end{cases} \\ B_{i,d}(u) &= \frac{u - t_i}{t_{i+d} - t_i} B_{i,d-1}(u) + \frac{t_{i+d+1} - u}{t_{i+d} - t_i} B_{i+1,d-1}(u) \end{aligned} \tag{4.22}$$

where in degree 0 is a flat line, degree 1 is a triangular hat function, degree 2 is a hump of 3 parabolic pieces and degree 3 is a hump of 4 cubic pieces, as shown in Fig. 4.10.



(a) The curve of basic function in degree 0 (b) Basic function curve for different degrees

Figure 4.10: The curves of B-spline basic functions

Therefore, the line function is

$$L_k(u) = \sum_{i=0}^{n_{cp}} \boldsymbol{g}_{m_k^i} B_{i,d}(u) \tag{4.23}$$

where uniformed $u \in [0, 1]$. To interpolate the corresponding image global position, we heavily sample the points in the range to find the closest point on the curve with respect to the position in $\widehat{\boldsymbol{x}}_k^c$, which is $\widehat{\boldsymbol{p}}_k^c$. The corresponding point on the curve is

$$j_k = \underset{j \in [0,1]}{\operatorname{argmin}}\, dist(L_k(j), \widehat{\boldsymbol{p}}_k^c) \tag{4.24}$$

where function *dist* calculates the least square distance between two points.

4.3.2 Orientation Interpolation from 3D B-spline Curve

Besides positions, the proposed method could also interpolate the orientation from the tangent vector $\boldsymbol{l}_{t_k}$ and binormal vector $\boldsymbol{l}_{b_k}$ derived from the first and second derivative of curve

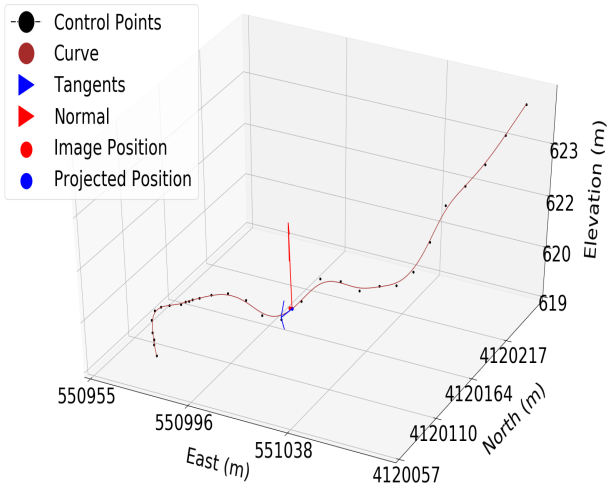
function.

$$\boldsymbol{l}_{t_k} = \frac{L'_k(j_k)}{\|L'_k(j_k)\|} = \begin{bmatrix} tx & ty & tz \end{bmatrix}^T \tag{4.25}$$

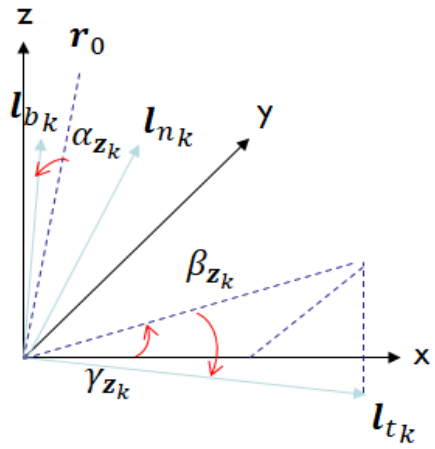
$$\boldsymbol{l}_{b_k} = \frac{L'_k(j_k) \times L''_k(j_k)}{\|L'_k(j_k) \times L''_k(j_k)\|} \tag{4.26}$$

As the z axis of vehicle frame should point up in the world frame, we correct the sign of binormal vector for the z axis of vehicle frame.

$$\boldsymbol{l}_{b_k} = -sign(\cos^{-1}(\boldsymbol{l}_{b_k}^T \cdot \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^T) - \frac{\pi}{2}) * \boldsymbol{l}_{b_k} \tag{4.27}$$



(a) Interpolate GPS measurement from the 3D curve constructed in the sliding window



(b) Orientation estimation

Figure 4.11: Interpolate angles of axises from 3D curve constructed by GPS points

As shown in Fig. 4.11, we can find the orientation angle around y axis by the following equation

$$\beta_{\boldsymbol{z}_k^c} = -\sin^{-1}(tz) \tag{4.28}$$

and the corresponding rotation matrix is

$$\boldsymbol{R}_{\beta_{\boldsymbol{z}_k^c}} = \begin{bmatrix} \cos(\beta_{\boldsymbol{z}_k^c}) & 0 & \sin(\beta_{\boldsymbol{z}_k^c}) \\ 0 & 1 & 0 \\ -\sin(\beta_{\boldsymbol{z}_k^c}) & 0 & \cos(\beta_{\boldsymbol{z}_k^c}) \end{bmatrix} \tag{4.29}$$

The orientation angle around z axis is

$$\gamma_{\boldsymbol{z}_k^c} = double(tx < 0) * sign(ty) * \pi + \tan^{-1}(\frac{ty}{tx}) \tag{4.30}$$

where *double* is the function that convert logic result to number. To find the orientation along x axis, we find it by the angle between binormal vector and the z axis after the rotation along y axis, $\boldsymbol{r}_0 = \boldsymbol{R}_{\beta_{\boldsymbol{z}_k^c}} \cdot \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^T$.

$$\alpha_{\boldsymbol{z}_k^c} = sign\left((\boldsymbol{r}_0 \times \boldsymbol{l}_{b_k}) \cdot \boldsymbol{l}_{t_k}\right) \cos^{-1}\left({\boldsymbol{l}_{b_k}}^T \cdot \boldsymbol{r}_0\right) \tag{4.31}$$

Then the position estimate from the GPS for each image is realigned as:

$$\boldsymbol{z}_k^c = \begin{bmatrix} L_k(j)^T & \alpha_{\boldsymbol{z}_k^c} & \beta_{\boldsymbol{z}_k^c} & \gamma_{\boldsymbol{z}_k^c} \end{bmatrix}^T \in \mathbb{R}^6 \tag{4.32}$$

4.4 Summary

The entire framework of the proposed AEKF-based camera global localization is summarized in Fig. 4.12. The proposed technique is built up based on the conventional EKF model with modifications in prediction, correction and adaptive covariance matrix with the details that demonstrated in this chapter. The initial state $\boldsymbol{x}_0$ is estimated from GPS readings in the sliding window. Also, the first m time steps residual $\boldsymbol{e}_{0:m-1}$ and innovation $\boldsymbol{d}_{0:m-1}$ are

initialized based on the fixed noise covariance, $\boldsymbol{W}_0$ and $\boldsymbol{Q}_0$.

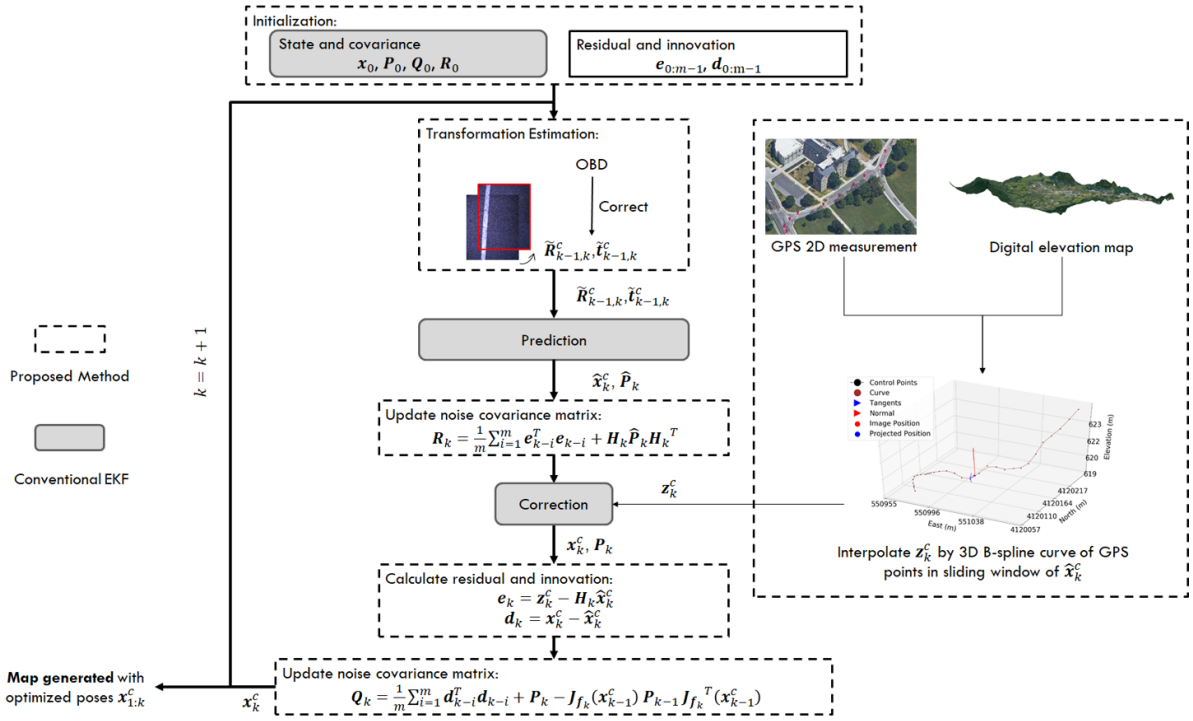


Figure 4.12: Proposed AEKF-based camera global localization framework

Chapter 5

3D Global Road Surface Mapping

The proposed global road surface mapping combines the local road surface mapping technique with camera global localization to avoid the accumulated error in vertical direction from vision. The overview of the proposed technique is shown in Fig. 5.1. The entire road

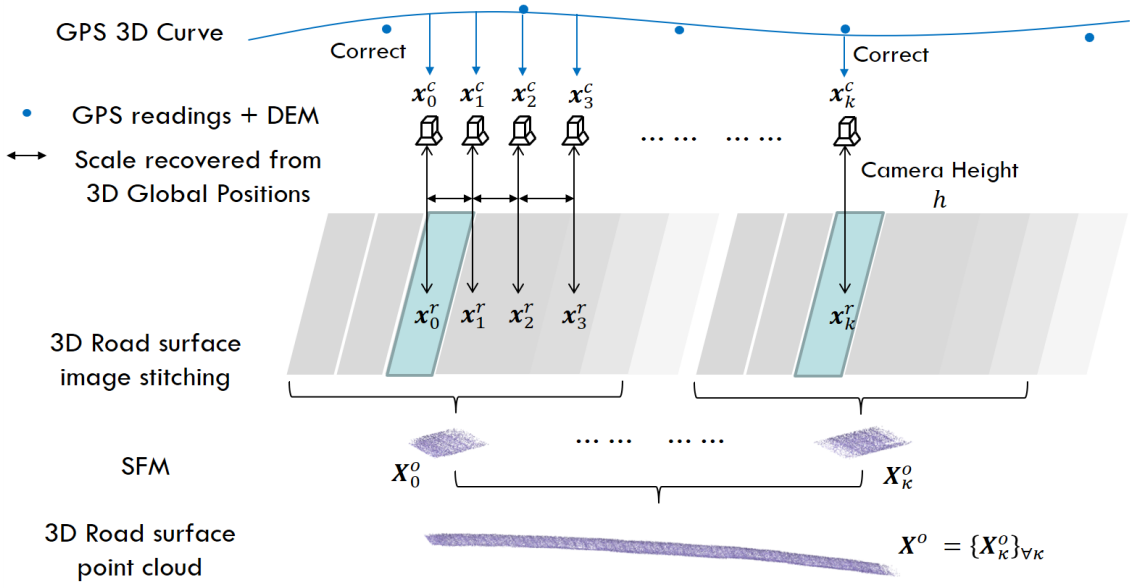


Figure 5.1: Overview of global road surface mapping

surface is generated with sequences of road surface segment. The road surface segment in global is generated with SFM which will be demonstrated in Sec. 5.1 and the final optimization and road assembly are described in Sec. 5.2.

5.1 Road Surface 3D Reconstruction from Multi-View

The past work mentioned that the error accumulated in vertical direction from vision is significant in multiple view road surface reconstruction. As the road surface is low-texture material comparing to buildings, trees and etc. Therefore, the proposed global road surface mapping technique choose to combining the local road surface mapping technique with camera global localization.

In the designed system, the same road segment can be covered in 3 to 5 images. Thus, to improve the accuracy of road surface reconstruction, the proposed method for local road surface mapping applied triangulation with multiple views and limited feature tracking. The global road map segment is generated from the triangulation of multiple views. In each road segment at step κ, there are m_v image is included for road surface 3D reconstruction. Also, for having road surface in global coordinate, the optimized global camera pose is used for rescale the road surface.

$$\boldsymbol{X}_{\kappa}^{o} = \boldsymbol{f}(\boldsymbol{u}_{m_v\kappa:m_v\kappa+m_v-1}, \boldsymbol{x}_{m_v\kappa:m_v\kappa+m_v-1}^{c}, \boldsymbol{K}) \tag{5.1}$$

For monocular vision 3D reconstruction method, the most common method is Structure from Motion (SFM) which has been described at the beginning. And as reviewed before, in the pure vision-based method in image localization, BA is applied to solve the scale problem and refined the position based on the distance in the first two image translation in unit, which is know as VisualSFM. In the proposed method, SFM is applied for generating 3D point cloud and scale is redefined by the optimized camera pose from the proposed AEKF-based camera global localization. The triangulation of multiple views is shown in Fig. 5.2. At time steps k, for each feature, there are corresponding features can be found in the next m_v views. All the features generated in m_v views will be the global road surface map at time

step κ, which is different frequency of camera.

As we all known that there are not significant features can be identified between road surface images. Therefore, for normal feature detection and match algorithm, it will also find some common features from two similar road surface that are captured from two different place. To avoid this problem, when we applied SFM for road surface 3D reconstruction, the feature tracking is only applied no more than from three previous time steps. Thus, the tracking feature existed over 4 frames will be rejected in this thesis. The interest point $\boldsymbol{x}^o_{k_j}$ for global road surface map is estimated based on the rescaled extrinsic matrix. The way to find the ideal world point position and transformation is the minimize the reprojection error.

$$\boldsymbol{X}^o_\kappa = \underset{\{\boldsymbol{x}^o_{k_j}\}_{\forall j}}{\operatorname{argmin}} \sum_{i=0}^{m_v-1} \sum_{j=1}^{n} dist\left(\boldsymbol{u}^j_{m_v\kappa+i}, \boldsymbol{M}_{m_v\kappa+i}\boldsymbol{x}^o_{k_j}\right)^2 \tag{5.2}$$

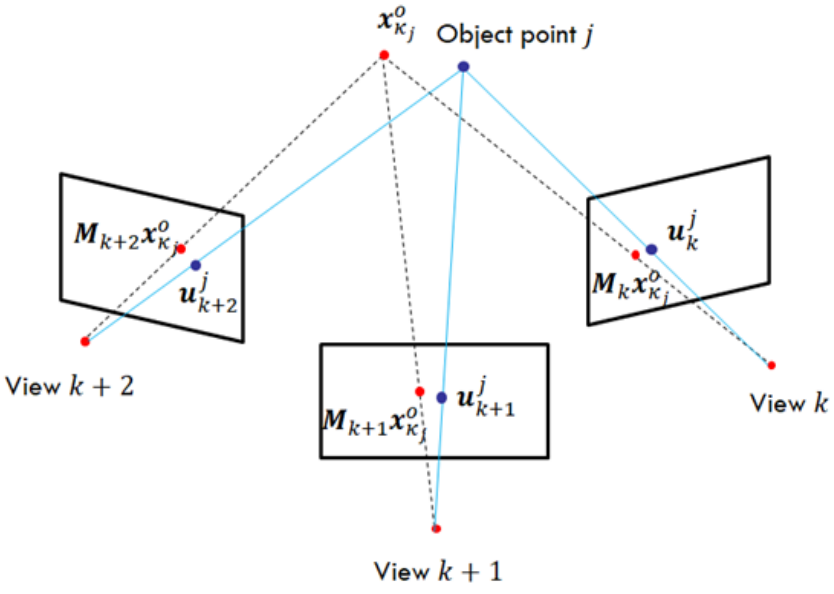


Figure 5.2: Triangulation from multiple views

5.2 Optimization with Bundle Adjustment

The reprojection error is defined by the distance between the observation of image features and estimated position from reprojection. It can be refined as an appropriate cost function, sum of squared errors, so that the optimize transformation and world points can be found by Levenberg-Marquardt algorithm [21].

The cost function for minimizing road surface mapping problem is

$$\Phi(\boldsymbol{M}_{m_v\kappa:m_v\kappa+m_v-1}, \boldsymbol{X}^o_\kappa) = \sum_{i=0}^{m_v-1}\sum_{j=1}^{n} dist\left(\boldsymbol{u}^j_{m_v\kappa+i}, \boldsymbol{M}_{m_v\kappa+i}\boldsymbol{x}^o_{k_j}\right)^2 \tag{5.3}$$

where $\boldsymbol{X}^o_\kappa = \{\boldsymbol{X}^o_{\kappa_j}\}_{\forall j}$. The cost function (sum of squared errors) is minimized with Bundle Adjustment (BA) to determine an optimal estimation of parameters, camera pose and road surface, and it reduces the inconsistency of locally camera motion estimation in global.

In this thesis, the existed technique Sparse BA is applied to solve the least-square problem, which is an efficient Levenberg-Marquardt (LM) based sparse implementation of BA. Finally, the camera pose and road surface will be optimized and the final global road surface map is generated by summing all global road surface segments.

$$\boldsymbol{X}^o = \{\boldsymbol{X}^o_\kappa\}_{\forall\kappa} \tag{5.4}$$

Chapter 6

Experiments and Results

This chapter provides the details of experiment and results analysis. The test platform, system design and test environment will be demonstrated in Sec. 6.2. And the parameters estimation, the evaluation of frame rate control and model performance analysis is covered in Sec. 6.3.

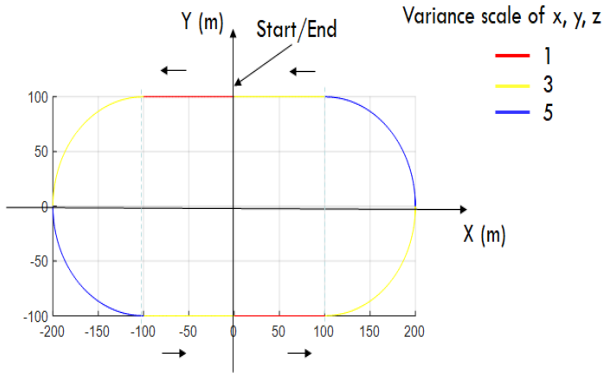
6.1 Simulation Experiments and Results

As the camera is set parallel to the ground surface, the global road surface mapping problem can be simplified as camera global localization. In the world coordinate, it is hard to find the real ground truth of global position. Therefore, an simulation world is established in this section to test the proposed AEKF model with the deployment of a sparse low-accuracy GPS for camera localization.

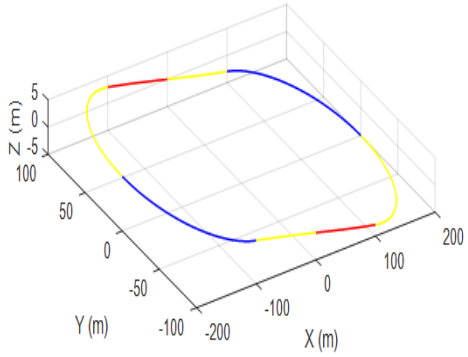
The designed simulation route is shown in Fig. 6.1 with both the straight lines and curves. The elevation is setup with function of time step t

$$z = 5 * \sin(\pi t/2000) \tag{6.1}$$

And the measurement variances are changed by scaling from initial variances in three different levels and indicated with different color along the route. The parameters are variances, time



(a) Designed simulation world with variance change



(b) Simulation world in 3D

Figure 6.1: Simulation world

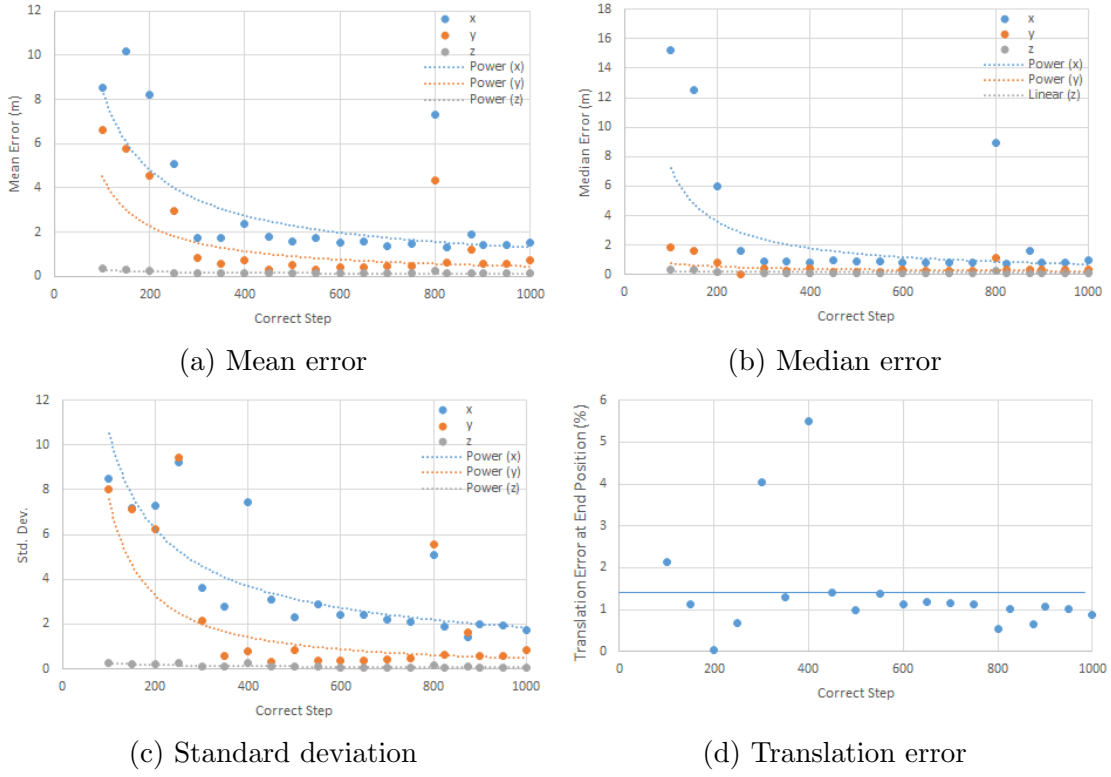
step size and correct step. The simulated GPS and image transformation measurements are generated randomly based on the ground truth and calculated variance. The noise matrix during the experiment is defined as

$$\boldsymbol{R}_0 = \begin{bmatrix} \sigma^2_{gx_0} & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma^2_{gy_0} & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2_{gz_0} & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2_{g\alpha_0} & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2_{g\beta_0} & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma^2_{g\gamma_0} \end{bmatrix} \quad \boldsymbol{Q}_0 = \begin{bmatrix} \sigma^2_{ix_0} & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma^2_{iy_0} & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2_{iz_0} & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2_{i\alpha_0} & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2_{i\beta_0} & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma^2_{i\gamma_0} \end{bmatrix} \tag{6.2}$$

With the similar method, the initial process noise matrix is found based on the error of image shift in 2D and rest of variances are assumed to be small as light change in small distance. And the initial error covariance also assumed to be small scale of identity matrix, $\boldsymbol{P}_0 = 0.001\boldsymbol{I}$. The parameter used in simulation world is shown in Table. 6.1. The ideal window size of covariances adjustment is not been decided. Therefore, in the simulation environment, multiple window size simulation is implemented and analyzed.

Table 6.1: Parameters used in the parametric studies

Parameters	**Values**
$\sigma^2_{gx_0}, \sigma^2_{gy_0}, \sigma^2_{gz_0}$	0.1
$\sigma^2_{g\alpha_0}, \sigma^2_{g\beta_0}, \sigma^2_{g\gamma_0}$	1
$\sigma^2_{ix_0}, \sigma^2_{iy_0}, \sigma^2_{iz_0}$	1e-05
$\sigma^2_{i\alpha_0}, \sigma^2_{i\beta_0}, \sigma^2_{i\gamma_0}$	0.01
Correct Steps m	100:50:1000
Unit travel distance (m)	0.25, $\pi/12$
Total length (m)	1028
Steps t	1:1:4000



(a) Mean error

(b) Median error

(c) Standard deviation

(d) Translation error

Figure 6.2: Error analysis of different window size for adaptive covariance matrix

As shown in Fig. 6.2, the ideal window size for adaptively adjusting covariance is studied. From the simulation results for different correct steps, the error tends to decrease as correct step increases, and keeps steady when the step size over 500. Therefore, parameter

m is set up as 600 so that the AEKF model adaptively adjust process and measurement covariances from the last 600 time steps. Then proposed method and conventional EKF method are compared with the ideal correct steps, and results are shown in Fig. 6.3.

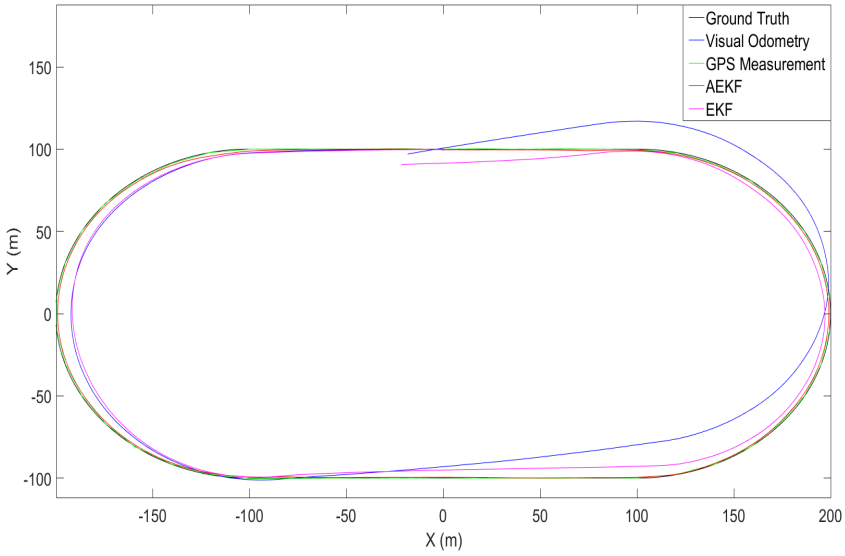
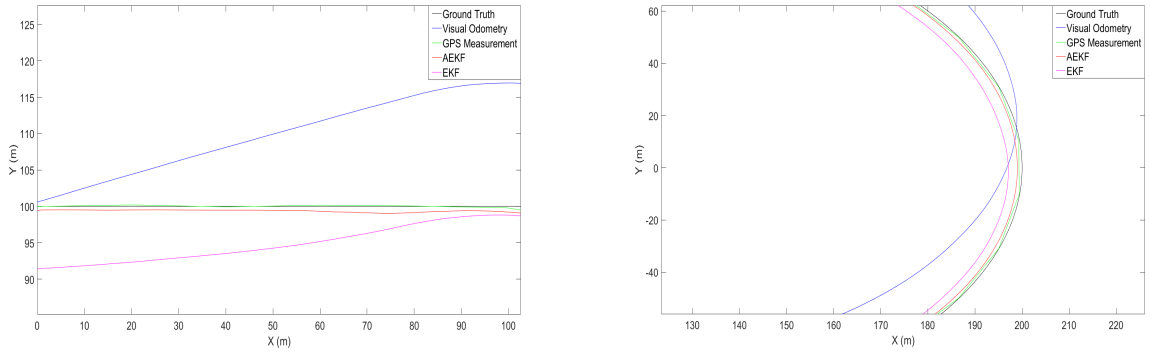


Figure 6.3: Simulation results with adaptive covariance window size 600

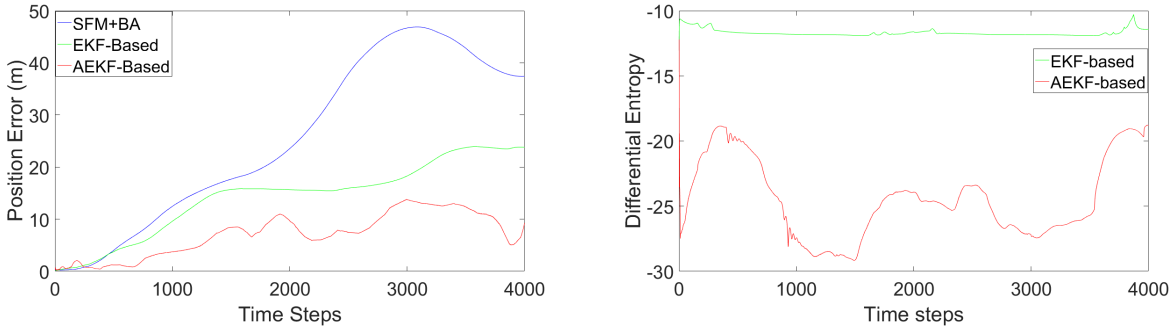
Both the classical EKF model and proposed AEKF-based model correct the position significantly from the accumulated error of visual odometry. The positions on the straight line at the end of a long sequence is shown in Fig. 6.4a and the positions on the curve line is shown in Fig. 6.4b. The corrected positions from the proposed AEKF-based model has close results to the ground truth and perform better than the classical EKF model. The mean position error of the proposed AEKF model has reduced 17.95% from the classical EKF-based model.



(a) Position results in straight line (b) Position results in curve line

Figure 6.4: Simulation results of a straight line and a curve line

Figure 6.5a quantitatively shows the transition of the position error with respect to time steps. The mean position error of the proposed AEKF-based technique has reduced 50.37% from the EKF-based technique. The differential entropy of the estimation by the AEKF-based technique is compared to the EKF-based technique in Figure 6.5b. While the differential entropy of the EKF-based technique remains high, it is seen that the proposed AEKF-based technique has low, varying values. This is due to the adaptive capability of the proposed technique.



(a) Position error vs. time steps (b) Differential entropy vs. time steps

Figure 6.5: Simulation result error analysis

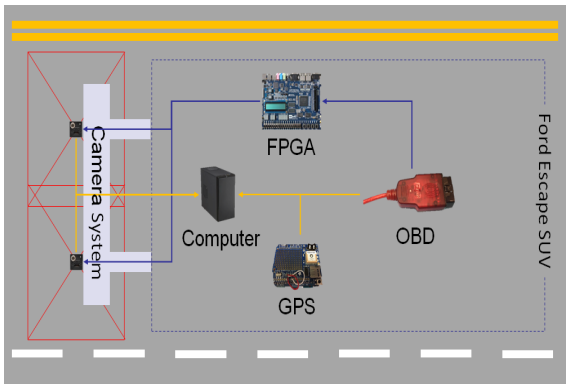
6.2 Field Experiments Setup

6.2.1 Hardware Architecture

The system consists of an camera system, a field programmable gate array (FPGA), a low-cost GPS, an on-board diagnostics (OBD) system as shown in Fig. 6.7. The stand-alone GPS module has update rate of 1 Hz and position accuracy of 3-5 m. The image capturing system we used in this paper refers to Hu and Furukawa's design [24] which has a pair of downward facing Complementary Metal-Oxide-Semiconductor (CMOS) cameras triggered and synchronized by a FPGA. The FPGA used in this research is Altera DE2-115. It captures high resolution (1280×1024) road surface images.

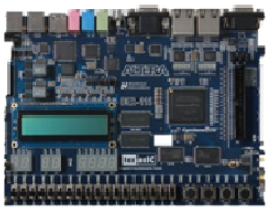
The entire system is assembled on the platform shown in Fig. 6.6. As the proposed method is applied for monocular vision, the designed system can put a pair of cameras are separated at a large distance to cover road width in the field of view, as shown in Fig. 6.6a. Our improved system controls the image frame rate based on real time vehicle speed from OBD port which presents in most current vehicles. The data flow of the designed system is shown in Fig. 6.6b. The designed system can maintain consistent overlapping region between consecutive images. This method efficiently covers road surface with enough images in different scenarios, such urban low speed, traffic light, highway, etc. Also, the steady translation in vehicle heading direction can be used to reject the mismatched features from moving shadows by adding motion constraint. The detail hardware specification and parameters is shown in Table. 6.2.

(a) Experiment Platform

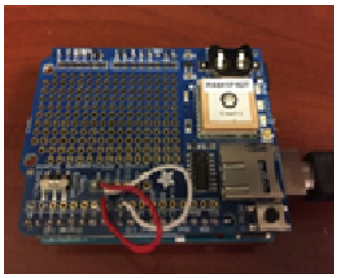


(b) Flowchart of system

Figure 6.6: System overview



(a) FPGA



(b) GPS

(c) OBD

Figure 6.7: Sensor overview

Table 6.2: Experimental hardware specifications

Parameter	Value
Image sensor	Point Grey Flea3
FOV	$56°09' \times 43°36'$
Image resolution	1280×1024 (pixel)
GPS sensor	Adafruit Ultimate GPS
GPS update rate	1Hz
GPS position accuracy	3-5m

6.2.2 Localization Environment

All experiment are implemented on the public road in Blacksburg, VA. The experiments included in this thesis are parts of the route in green shown in Fig. 6.8. The weather during test includes both sunny and cloudy so that there are different light condition in road surface images. There are two main experiments included in this thesis. One is a long experiments consists of local and highway with 6.9 km. The other experiment consists of multiple short routes in both local and highway.

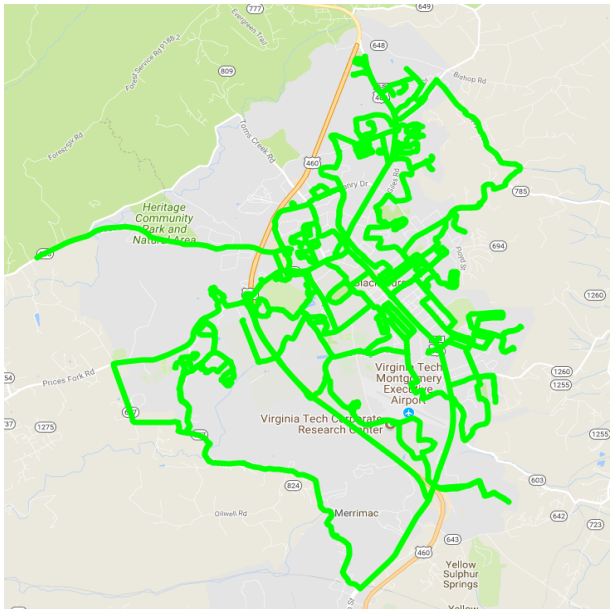


Figure 6.8: Localization environment

6.3 Field Experiment Results and Discussion

6.3.1 Camera Calibration Results

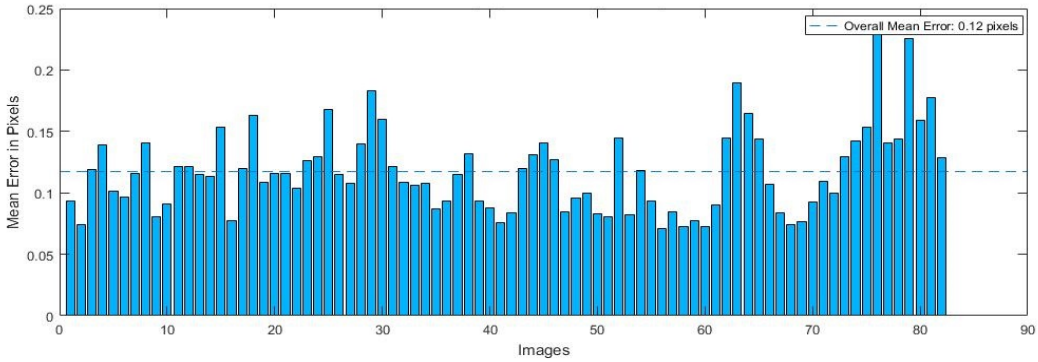
Before the experiments and results analysis, the camera needs to be calibrated to remove image distortion and measure the camera height based on the average extrinsic matrix. The camera calibration results for experiments are shown in Table. 6.3.

Table 6.3: Camera calibration results of experiments

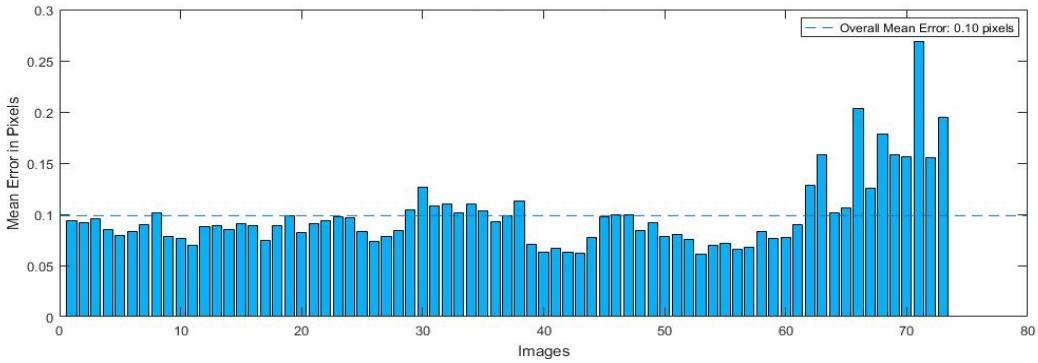
Experiments	**Parameters**	**x**	**y**
Exp. 1	Focal length f (pixels)	1341.4859 ± 0.9894	1341.3766 ± 0.9844
	Principal point c (pixels)	634.5761 ± 0.3703	514.3156 ± 0.4384
	skew s	-0.6612 ± 0.0548	
	Radial distortion d_r	-0.1982 ± 0.0005	0.1818 ± 0.0009
	Tangential distortion d_t	-0.0023 ± 0.0001	0.0001 ± 0.0001
Exp. 2	Focal length f (pixels)	1291.1900 ± 0.5876	1290.1659 ± 0.5792
	Principal point c (pixels)	623.7719 ± 0.3620	522.0138 ± 0.3747
	skew s	2.1792 ± 0.0514	
	Radial distortion d_r	-0.2077 ± 4.1090e-04	0.1483 ± 6.6660e-04
	Tangential distortion d_t	-0.0016 ± 4.9911e-05	6.5122e-05 ± 4.5862e-05

A reprojection error is the distance between the pattern keypoints detected in a calibration image and the corresponding world points projected into the same image. The overall mean reprojection error for both experiments are good, which is 0.1 pixels.

With having the camera calibrated, the camera height h is 1.5202 m for Exp.1 and 1.4510 m for Exp.2. Based on the FOV in Table. 6.2, the ratio of image shift, meter to pixel, is 0.011876 for Exp.1 and 0.001134 for Exp.2.



(a) Mean reprojection error of Exp.1



(b) Mean reprojection error of Exp.2

Figure 6.9: Mean reprojection error of each calibration image

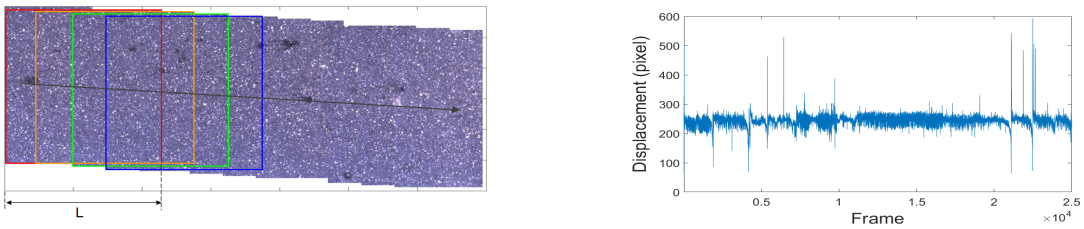
6.3.2 Performance Evaluation of Frame Rate Control

To keep enough overlapping region for good features matching, the frame rate fr is defined by

$$fr = \frac{Vn}{3.6l} \tag{6.3}$$

where V is the vehicle speed in km/h and n is the number of image that cover the same region in l. An example is shown in Fig. 6.10. The number of image covered in l is set as 4 in the experiments.

To evaluate the performance of control, the translation between I_k and I_{k-1} in pixels with the proposed method is calculated. In results, as shown in Fig. 6.10, the average



(a) Illustration of image overlapping (b) Frame rate evaluation based on image shift

Figure 6.10: Dynamically frame rate control based on vehicle speed and the evaluation of overlapping

translation with respect to vehicle moving direction is 235 pixels, which is around quarter of image length along the vehicle moving direction as designed. The lower bound of pixel displacement is around 200 which used as a threshold of motion constraint in proposed method. As shown in the figure, there are some peaks and valleys which happened when the vehicle braked or accelerated in a short time.

6.3.3 Field Experiments and Results

The field experiments are selected from the experiments, which consist of multiple routes, in Fig. 6.11. The selected experiments are consists of two main experiments with different camera calibration. One is a long route in sunny day that up to 6.9 km. The other one, Exp.2, consists of multiple routes in different places, times and weather as shown in Table. 6.4. And for each routes, it consists of different road sequences that are indicated with letters. For example, Exp. 2.1.f means that experiment uses the second camera calibration results and it is the f road sequence in route 1.

There are two methods to evaluate the proposed method. For localization, the common method to evaluate the results in position accuracy. For global localization evaluation, it is hard to find the ground truth. In this research work, the satellite imagery is chosen as ground truth by comparing the road markings from the stitched road map. The selected

Table 6.4: Experiments on the public road of Blacksburg

Exp. #	1	2.1.f	2.2.b	2.3.a
Weather	Sunny	Cloudy	Partly Cloudy	Partly Cloudy
Time	12:40 PM	10:30 AM	1:40 PM	6:20 AM
Length (mi)	4.29	2.04	1.30	2.10
Image #	25383	11685	6983	8826
Start image #	1	78094	21371	1
End image #	25383	89778	28353	8826

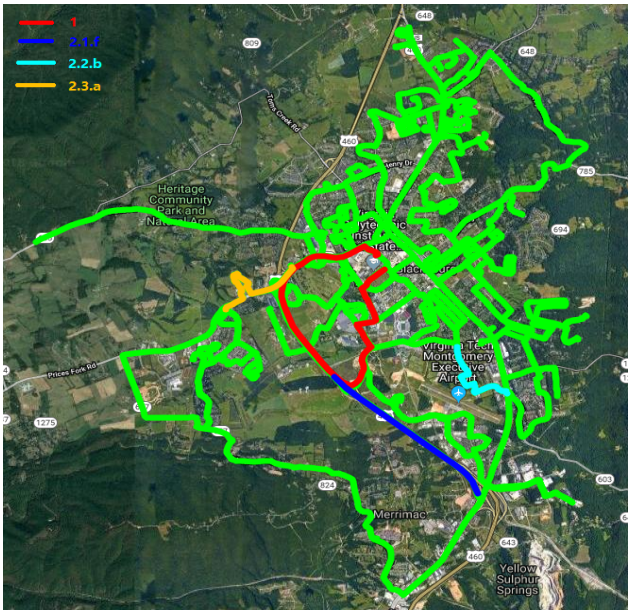
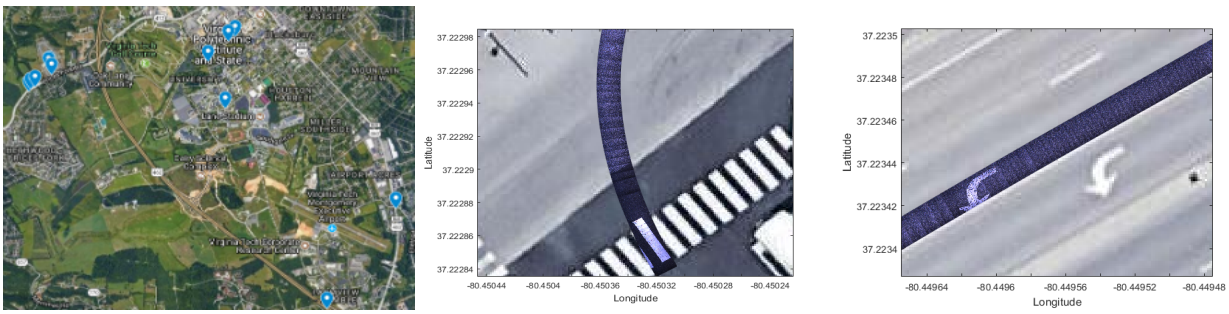


Figure 6.11: Selected routes in field experiments

routes contains both local and highway, and local has more road markings and selected as position accuracy analysis. The example of selected road markings are shown in Fig. 6.12a with blue point mark. Some points has good location results as Fig, 6.12b and some points have bad results as Fig. 6.12c.



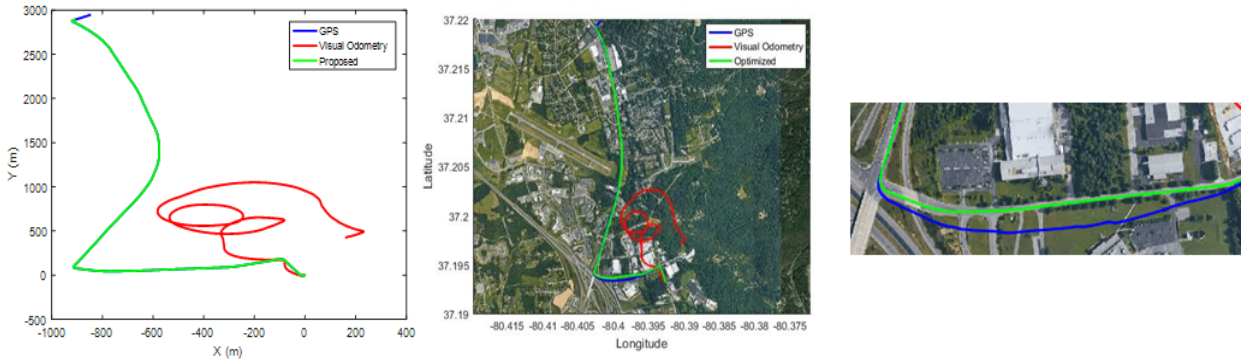
(a) Selected points (b) Good position (c) Bad position

Figure 6.12: Selected road markings for position accuracy analysis

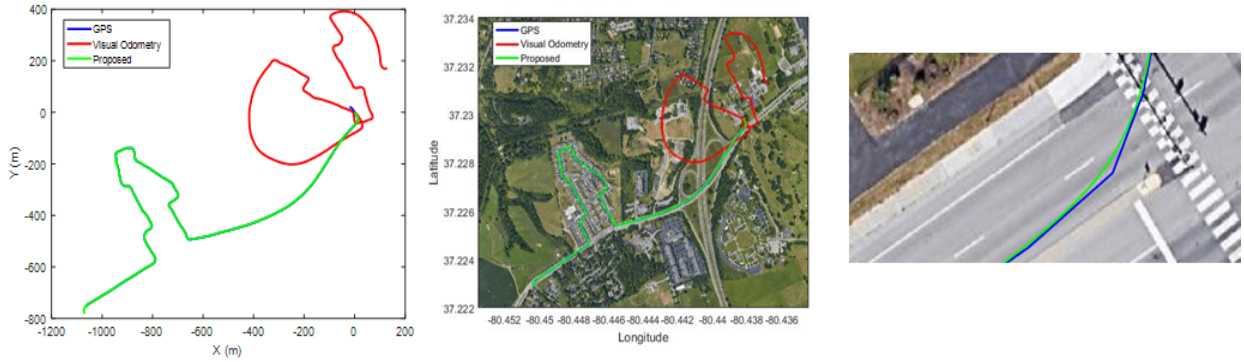
For building the road surface map, images are stitched based on calculated scale and image transformation. Therefore, the other main error source is the accumulated error from monocular visual odometry. The errors of the selected road markings in x, y, z is calculated by comparing the position of the corresponding road markings. Positions are converted from latitude and longitude to UTM coordinates, and the errors are tabulated in Table. 6.5. Two more corrected results in both local coordinate and global coordinate are shown in Fig. 6.13, which corrected the accumulated error from visual odometry with GPS measurement. The zoomed in results are illustrated on the right side of figures.

Table 6.5: Position error analysis of experiments

Experiments	**Direction**	**Mean (m)**	**Median (m)**	**Std. Dev.**
1	x	2.16	2.14	1.30
	y	3.32	1.78	3.28
	z	0.20	0.19	0.12
2.1.f	x	2.42	2.42	0.59
	y	1.24	1.24	0.94
	z	8.58	8.58	12.09
2.2.b	x	2.58	2.28	2.31
	y	3.61	4.14	2.43
	z	0.64	0.20	0.92
2.3.a	x	0.90	0.32	1.30
	y	2.27	1.63	2.45
	z	0.14	0.14	0.65
Total	x	2.24	2.00	1.76
	y	2.90	1.79	2.57
	z	0.38	0.18	0.65



(a) Localization results 1

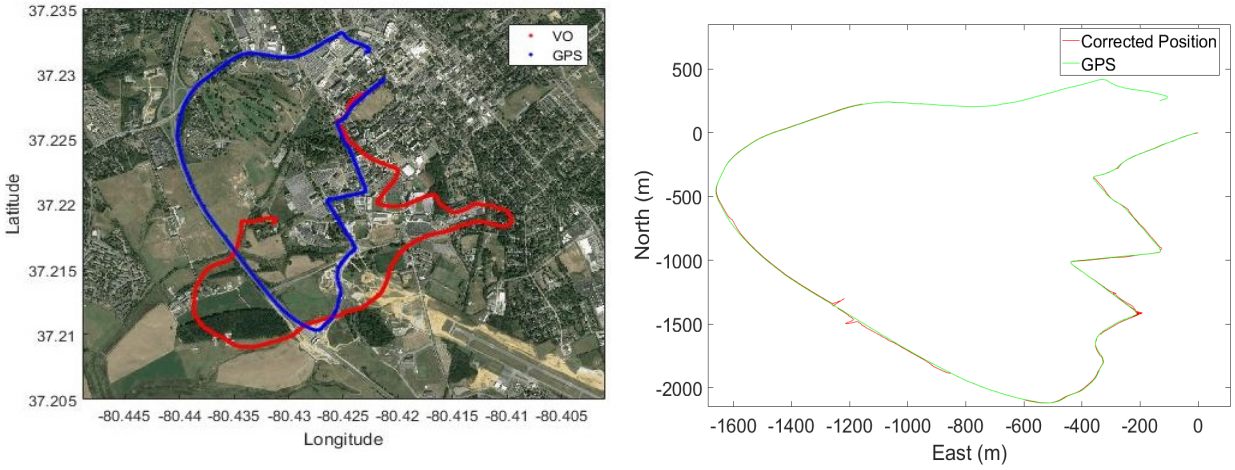


(b) Localization results 2

Figure 6.13: Localization results

The other method to evaluated the results is for visual odometry. The accumulated error from visual odometry affect the results, especially for long road sequence. The longest route in selected experiment has about 6.9 km and consists of both urban, which has more sharp turning curves, and highway, which has smooth straight lines, located in Blacksburg, Virginia, as shown in Fig. 6.14a. Low-precision GPS shows local inaccuracy but does not suffer from accumulation of errors. In long range, accumulation of errors of visual odometry results in the end position being far off the destination. The corrected trajectory with our estimation method is shown in Fig. 6.14b.

Result shows that the visual odometry has been corrected along with GPS position. Also, it is obvious to find that errors around turns are more than straight roads, and accumu-



(a) Comparison between the trajectory estimated by vision-only and raw GPS in world coordinates

(b) Corrected camera positions and raw GPS data

Figure 6.14: Long distance camera localization result

lated error caused drift in long range. Some positions have large drift since the orientation of image is not corrected efficiently, which leading to accumulation of errors, but they will be draw back from position measurements.

To validate the accumulated error from visual odometry, the error is calculated in the same way as in the KITTI evaluation that comparing the visual result with the ground truth at the endpoint of the sequences and calculate the translation errors as percentages of the traveled distances. The images are reprojected back to satellite imagery to compare the position of road markings. The total frames in this experiment is 25383, and the route

Table 6.6: Compare translation error over travel distance

Distance (m)	**Frames**	**Translation error %**
491.3	1864	1.70
1156.9	4333	0.85
1478.2	5517	5.95
5836.0	22782	6.85
6902.9	25227	7.97

starts and ends in urban area. The translation error is found by tracking road markings in the image sequences, and most of comparison points are in the urban area. From the results listed in Table 6.6, we find that the translation error increases over time. More errors occur on sharp turns of urban road, while less errors on highways. The best result of global position estimation obtained is in shorter distance, 1.157 km, with translation error 0.85%. For the longest distance, 6.9 km, it has the translation error 7.97%. With the same method, the translation error is also evaluated for other field experiments and the results are shown in Table. 6.7.

Table 6.7: Translation error over travel distance for field experiments

Experiment	**Distance (km)**	**Translation error** %
1	6.90	7.97
2.1.f	3.28	2.29
2.2.b	2.09	5.44
2.3.a	3.38	0.017

The best results in the field experiments are the last one that the translation error is 0.017% over 3.38 km. Also, the translation errors are vary in different experiment. For example, in the first experiment, as the data is collected in the sunny day, the vehicle moving shadow has more effects on the results than other experiments so that the error is much higher.

Also, the results are compared with some other previous works. Although there is no relevant road surface mapping data from previous literatures, there are some works uses the similar techniques that use downward vision as visual odometry for localization. The comparison is shown in Table. 6.8, and the proposed method achieves better results than previous techniques.

Table 6.8: Comparison of Translation error

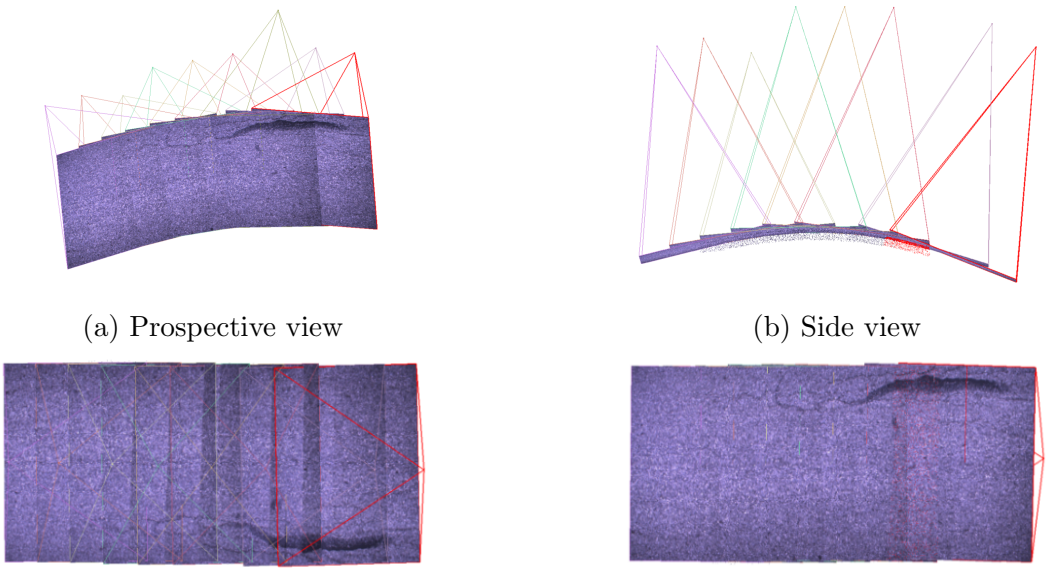
Reference	**DOF**	**Method**	**Distance**	**Translation error, end position**
Ericson and Astrand (2008) [16]	2	stereo DW	0.6m	0.7%
Piyathilaka and Munasinghe (2011) [32]	3	mono DW, optical flow	35m	17%
Ericson and Astrand (2017) [15]	6	stereo DW, ORB-SLAM	1.8-3.1km	2.63%-15.0%
Proposed Method	6	mono DW, GPS, BA, AEKF	2.09-6.9km	0.017%-7.97%

6.3.4 Road Surface Map Visualization

Road Surface Image 3D Stitching

The proposed method can be used to generate the road surface map in arbitrary length. The above section shows the camera localization results in different distance. With having the optimized camera pose, the road surface images can be reprojected in to global coordinate as earth ground surface.

For having a clear view of the road surface details, the short segments of the results are selected for results visualization. The image stitching result shown in Fig. 6.15 is a short road segment about 2 meters. The image stitched in global coordinate is shown in Fig. 6.16 and Fig. 6.17. The example of the high resolution road surface images projected over satellite image with estimated global positions and orientations. Red points indicate the corresponding position on the satellite image. Also, the pedestrian walkway and the edge of road marking are closely matched with the satellite image.



(a) Prospective view (b) Side view

(c) Top view (d) Bottom view

Figure 6.15: Road surface image stitching in 3D

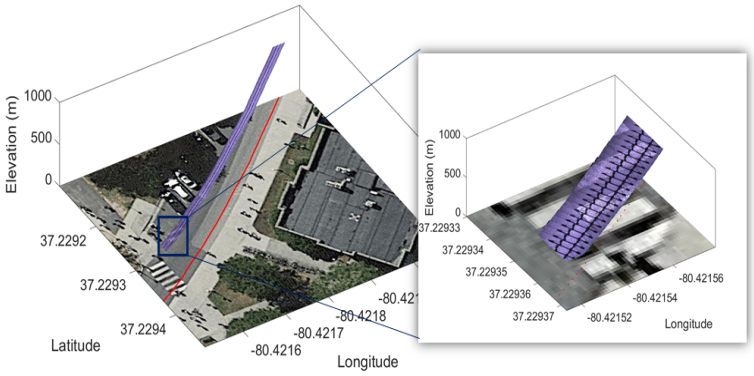


Figure 6.16: Images projected over the satellite image

Road Surface Point Cloud 3D Stitching

With the corrected positions, the results can be shown not only as stitching images, but also as stitching point cloud for 3D road surface reconstruction. The results from the proposed method are compared with the classical 3D reconstruction method, VisualSFM, which is a pure vision-based method.

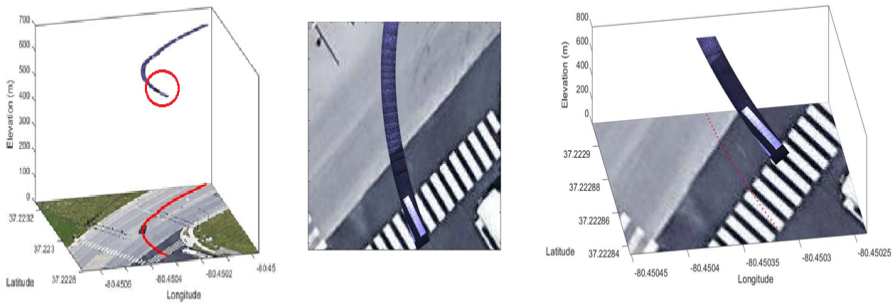
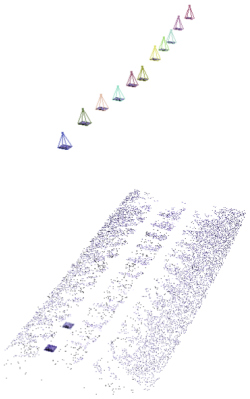
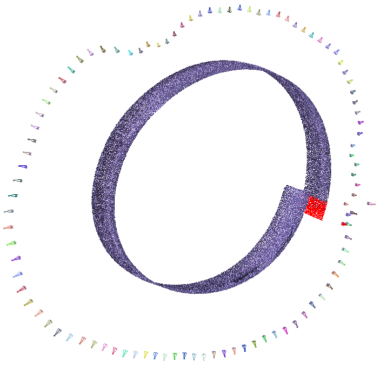


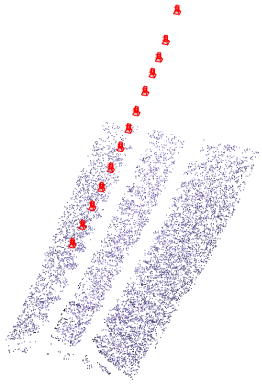
Figure 6.17: Global road surface image stitching results in both 2D and 3D



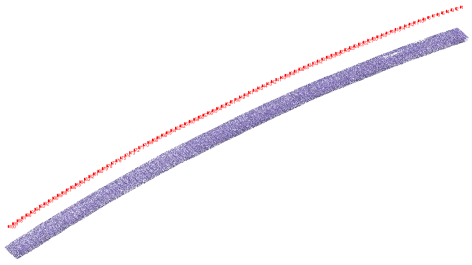
(a) VisualSFM in short distance



(b) VisualSFM in long distance



(c) Proposed method result in short distance



(d) Proposed method result in long distance

Figure 6.18: Comparison of the results of VisualSFM method and vision sensor fusion method for low texture environment 3D point cloud reconstruction

The point clouds and the camera position is shown in each sub-figures. As road surface images has limited significant features, there are wrong position estimation in Fig.6.18a and Fig. 6.18b. In the short segment, the position at the end has obvious wrong transformation, but the point clouds still keep in good performance. For the long road segment, the accumulated error from visual odometry in low texture environment causes the bad point cloud stitching and unreasonable results. With the proposed multiple view triangulation and BA method, the road surface point clouds are shown as straight curve.

Also, as shown in Fig. 6.19, the stitching results can also shown the road defects, color information and etc. for many other applications. From the side view of stitching results, the pothole is clearly shown in the change of road profiles and the color information is also saved in point clouds. With stitching road surface point cloud in global, the 3D road surface mapping with geolocation is shown in Fig. 6.20 and zoomed in results in Fig. 6.21.

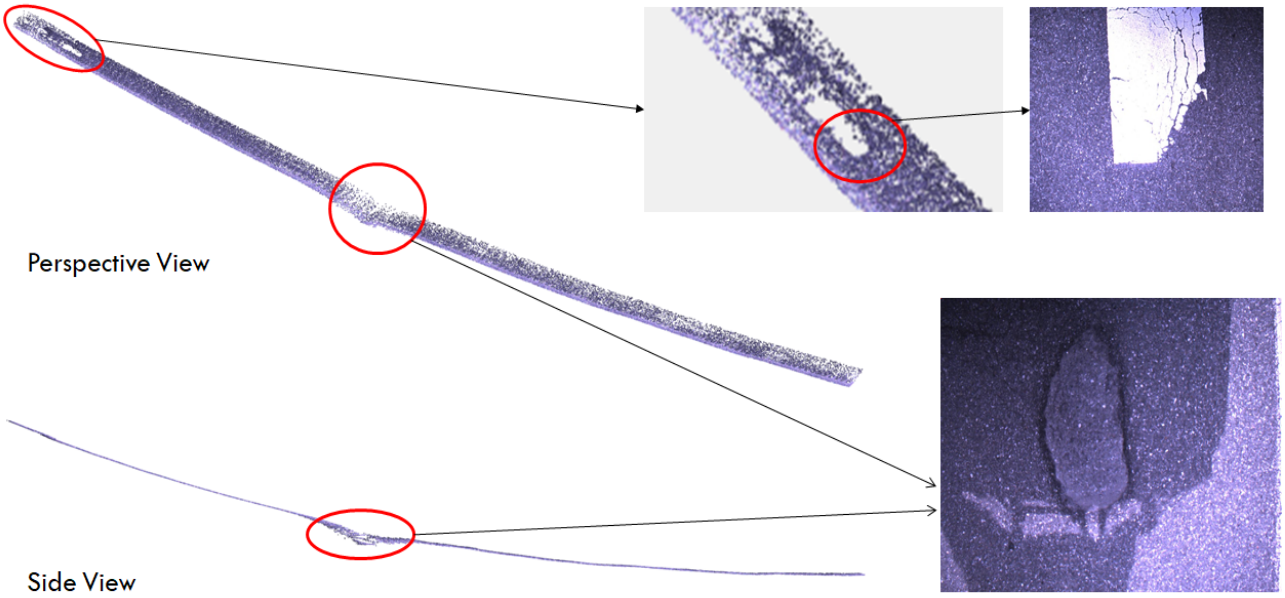


Figure 6.19: Road surface 3D point cloud stitching results

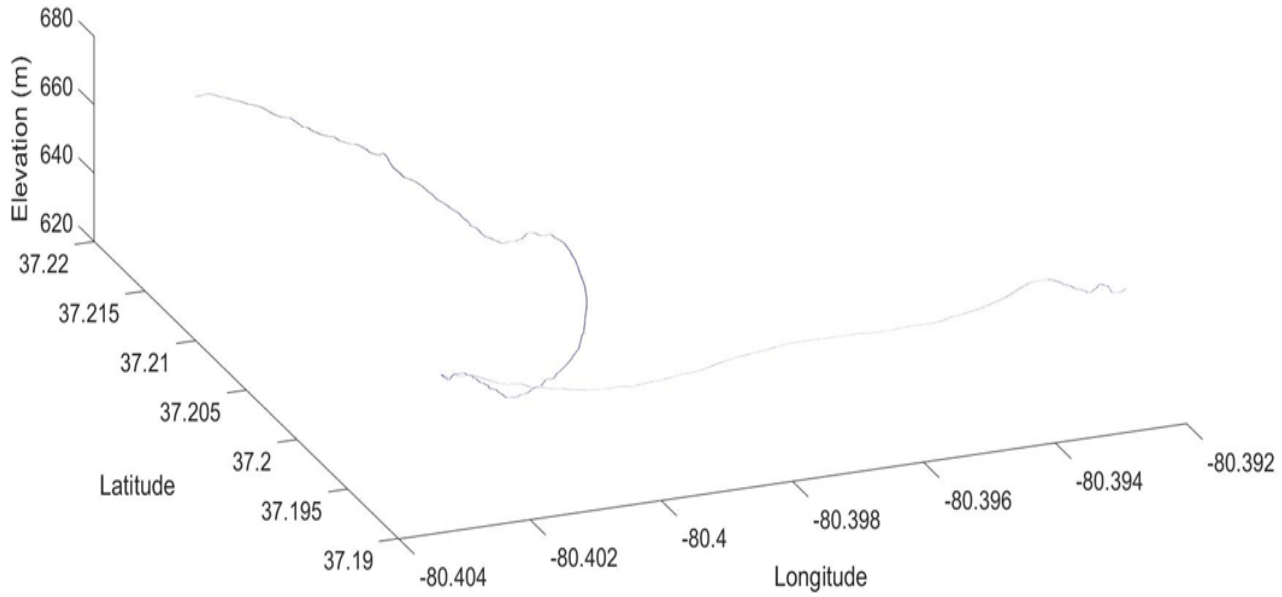


Figure 6.20: Road surface 3D point cloud stitching in global

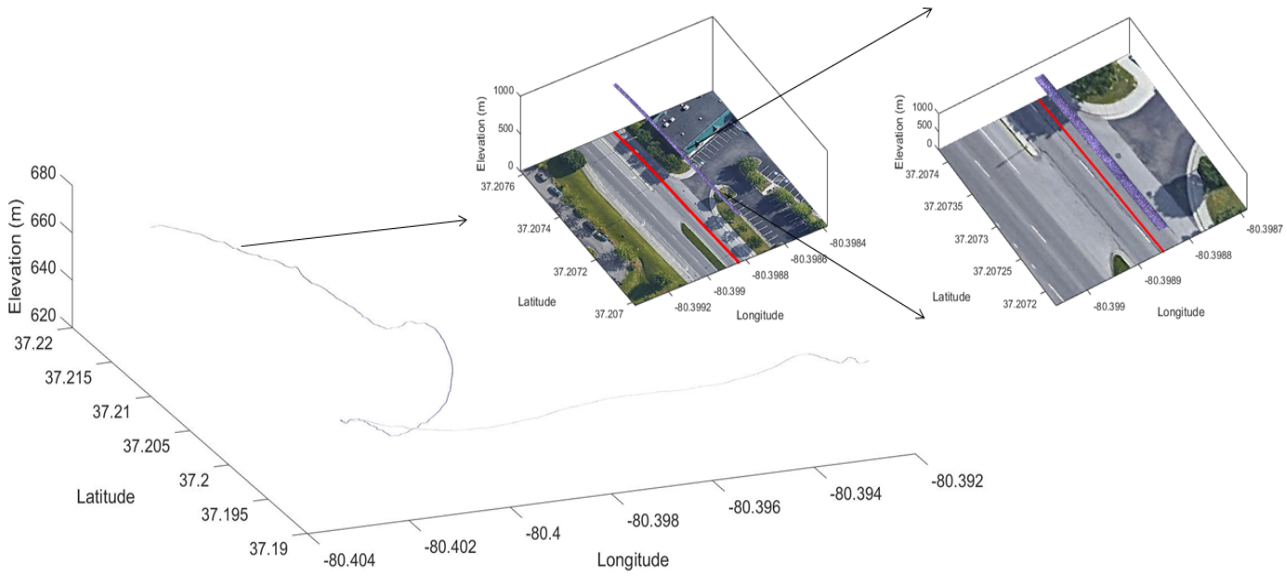


Figure 6.21: Globally reconstructed 3D road surface

Chapter 7

Conclusions and Future Work

7.1 Conclusion

This thesis has presented a simultaneous 3D mapping and geolocation of road surface technique that combines local road surface mapping and global camera localization. The vision-based road surface mapping system with a ground-facing camera was chosen in this research, which can collect rich road surface information. The system was improved that the camera frame rate is dynamically controlled by vehicle speed read from on-board diagnostics (OBD) for collecting continuous data.

The purpose of this research is to generate 3D road surface map with geolocation, which includes road profiles, such as elevation and orientation change in world coordinate. The global 3D road surface mapping was designed with combining local road surface mapping and global camera localization to solve the potential issues in low-texture environment. The local road surface was generated by structure from motion (SFM) with multiple views. The global camera localization was implemented with proposed Adaptive Extended Kalman Filter (AEKF)-based 3D global localization using image shift as prior. And the camera pose was corrected with the sparse low-precision Global Positioning System (GPS) data and digital elevation map (DEM). The optimized camera positions can help to rescale the local road surface map, and then the Bundle Adjustment (BA) was applied to reduce the projection errors. The final 3D road surface map with geolocation was generated by combining both

local road surface mapping and global localization results.

The final results was tested by camera localization in the simulation world. The parametric of AEKF was studied and the performance of the proposed AEKF-based camera global localization was compared with conventional EKF in simulation. The ideal correct steps is 600 and the proposed method reduced 50% from the conventional EKF-based model. In the field experiment, the position accuracy in each direction is 2.24m, 2.9m and 0.38m. For evaluating road surface mapping from vision, the translation error at the end position is calculated. The proposed method has the translation error 0.017%-7.97% over distance 2.09km-6.9km, which has achieved better results in longer distance than previous literatures.

7.2 Future Work

The future work can be conducted in two directions: improve image pose accuracy and extend application of high resolution infrastructure data. Under future work, the system can be improved by adding more sensors for localization, such as IMU to improve the accuracy on position and orientation, front view camera to localize respect to lane markings. As the camera is attached on the moving vehicle, getting more information of vehicle dynamics can help to predict the camera motion. Also, by adjusting the camera height, the camera field view can be increased to over the width of the road so that more road markings can be included in images. Then localization accuracy can be improved by aligning road marking positions in collected images with their position in satellite images or existed road surface map. As the limitation of FOV, the matched pattern can be found around the estimated trajectory as shown in Fig. 7.1. This will efficiently correct the position, especially in urban area with many road markings.

As the road surface has less objects than the images from forward-facing cameras,

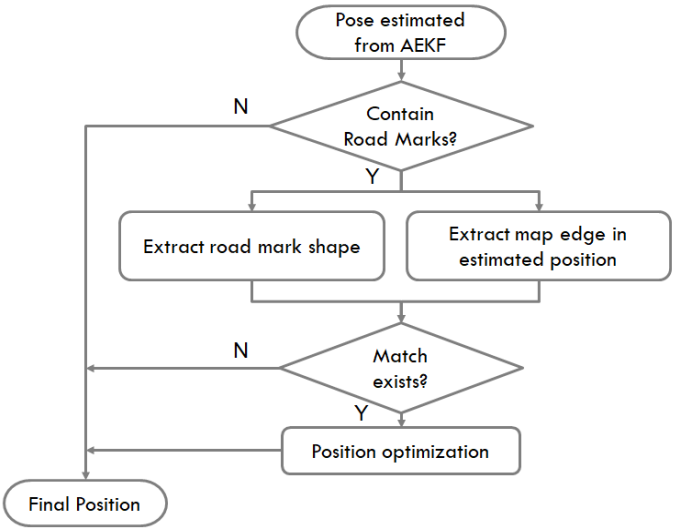


Figure 7.1: Flowchart of future work idea about improving position accuracy by map-based method

we can simply detect if the image has road markings by color as shown in Fig. 7.2. From the normal distribution of pixel value for each tonal value, we can simply classified road markings. As shown in the second row of Fig. 7.2, the image with road markings has a pulse around 250, which indicates the white color. Also, as doing off-line process, we can also apply other more computational and accurate methods to classify out the image contains road markings.

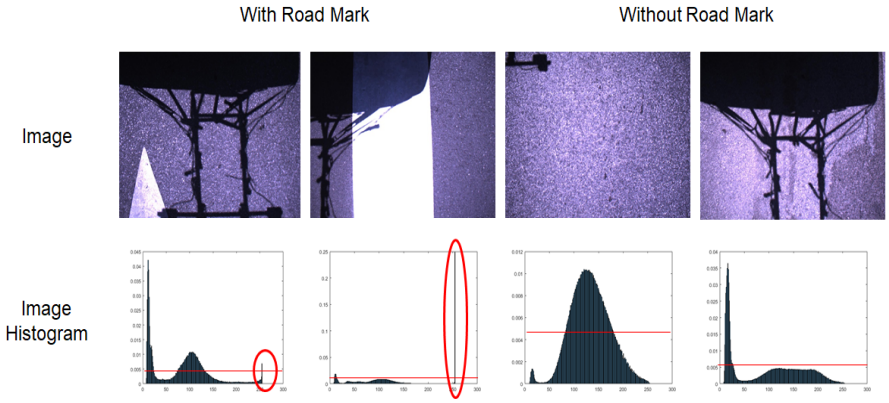


Figure 7.2: Detect if image has road markings by color

For the detected image, we can extract the shape of road markings from image for matching, as shown in Fig. 7.3. After converting the image to binary and remove noise by median filter, we extracted the road marking shape by tracking the boundaries. With having estimated global position for the selected image, we can extract an image in the region and detect edges for matching with the shape of the road marking.

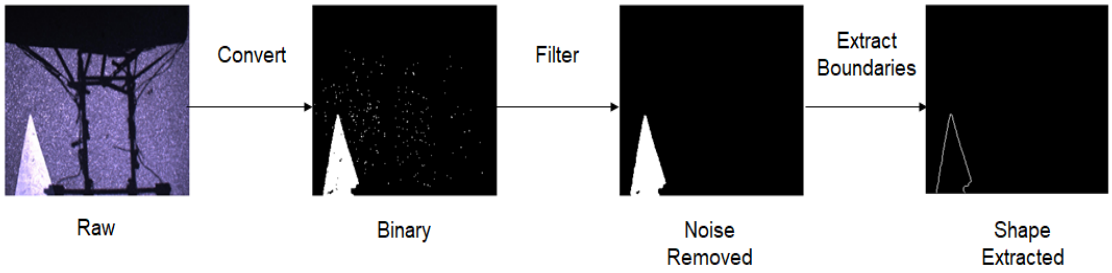


Figure 7.3: Extraction of the road marking shapes

The other direction of future work is the application of high resolution infrastructure data. As more and more intelligent system designed for improving driving safety, providing the road surface condition through Vehicle-to-Infrastructure (V2I) and Vehicle-to-Vehicle (V2V) can help vehicle or driver get more road information ahead, which can increase the driving safety and comfort.

Bibliography

[1] Table 1-37: U.s. passenger-miles. URL https://www.bts.gov/archive/publications/national_transportation_statistics/2006/table_01_37.

[2] *Rough roads ahead: fix them now or pay for it later*. AASHTO, 2009.

[3] *Bumpy roads ahead: Americas roughest rides and strategies to make our roads smoother*. TRIP, 2013.

[4] M. Agrawal and K. Konolige. Real-time localization in outdoor environments using stereo vision and inexpensive gps. *18th International Conference on Pattern Recognition (ICPR06)*, 2006. doi: 10.1109/icpr.2006.962.

[5] Pablo F Alcantarilla, Luis M Bergasa, and Frank Dellaert. Visual odometry priors for robust ekf-slam. *2010 IEEE International Conference on Robotics and Automation*, 2010. doi: 10.1109/robot.2010.5509272.

[6] Ahmad Alhasan, Kyle Younkin, and David J White. *Comparison of Roadway Roughness Derived from LIDAR and SFM 3D Point Clouds*. URL http://www.intrans.iastate.edu/research/documents/research-reports/roadway_roughness_w_cvr.pdf.

[7] Mohammad O. A. Aqel, Mohammad H. Marhaban, M. Iqbal Saripan, Napsiah Bt. Ismail, and Asem Khmag. Optimal configuration of a downward-facing monocular camera for visual odometry. *Indian Journal of Science and Technology*, 8(32), Oct 2016. doi: 10.17485/ijst/2015/v8i32/92101.

[8] Vitor B. Azevedo, Alberto F. De Souza, Lucas P. Veronese, Claudine Badue, and Mariella Berger. Real-time road surface mapping using stereo matching, v-disparity

and machine learning. *The 2013 International Joint Conference on Neural Networks (IJCNN)*, 2013. doi: 10.1109/ijcnn.2013.6707066.

[9] Sylvie Chambon and Jean-Marc Moliard. Automatic road pavement assessment with image processing: Review and comparison. *International Journal of Geophysics*, 2011: 120, 2011. doi: 10.1155/2011/989354.

[10] Xiao Chen, Weidong Hu, Lefeng Zhang, Zhiguang Shi, and Maisi Li. Integration of low-cost gnss and monocular cameras for simultaneous localization and mapping. *Sensors*, 18(7):2193, Jul 2018. doi: 10.3390/s18072193.

[11] Hyukdoo Choi, Dong Yeop Kim, Jae Pil Hwang, Chang-Woo Park, and Euntai Kim. Efficient simultaneous localization and mapping based on ceiling-view: Ceiling boundary feature map approach. *Advanced Robotics*, 26(5-6):653671, 2012. doi: 10.1163/156855311x617542.

[12] Weidong Ding, Jinling Wang, Chris Rizos, and Doug Kinlyside. Improving adaptive kalman estimation in gps/ins integration. *Journal of Navigation*, 60(03):517, Sep 2007. doi: 10.1017/s0373463307004316.

[13] D.w. Eggert, A. Lorusso, and R.b. Fisher. Estimating 3-d rigid body transformations: a comparison of four major algorithms. *Machine Vision and Applications*, 9(5-6):272290, Jan 1997. doi: 10.1007/s001380050048.

[14] Khalid L.a. El-Ashmawy. Investigation of the accuracy of google earth elevation data. *Artificial Satellites*, 51(3), Jan 2016. doi: 10.1515/arsa-2016-0008.

[15] Stefan Ericson. *Vision-Based Perception for Localization of Autonomous Agricultural Robots*. PhD thesis, 2017.

[16] Stefan Ericson and Astrand Bjorn. Visual odometry system for agricultural field robots. 2173, 10 2008.

[17] Rui Fan, Xiao Ai, and Naim Dahnoun. Road surface 3d reconstruction based on dense subpixel disparity map estimation. *IEEE Transactions on Image Processing*, 27(6): 30253035, 2018. doi: 10.1109/tip.2018.2808770.

[18] Duncan P. Frost, Olaf Kahler, and David W. Murray. Object-aware bundle adjustment for correcting monocular scale drift. *2016 IEEE International Conference on Robotics and Automation (ICRA)*, 2016. doi: 10.1109/icra.2016.7487680.

[19] Yang Gao, Ruofei Zhong, Tao Tang, Liuzhao Wang, and Xianlin Liu. Automatic extraction of pavement markings on streets from point cloud data of mobile lidar. *Measurement Science and Technology*, 28(8):085203, 2017. doi: 10.1088/1361-6501/aa76a3.

[20] Miguel Gaviln, David Balcones, Oscar Marcos, David F. Llorca, Miguel A. Sotelo, Ignacio Parra, Manuel Ocaa, Pedro Aliseda, Pedro Yarza, Alejandro Amrola, and et al. Adaptive road crack detection system by pavement classification. *Sensors*, 11(10): 96289657, Dec 2011. doi: 10.3390/s111009628.

[21] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge University Press, 2017.

[22] J. Heikkila and O. Silven. A four-step camera calibration procedure with implicit image correction. *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. doi: 10.1109/cvpr.1997.609468.

[23] C. Hide, T. Moore, and M. Smith. Adaptive kalman filtering algorithms for integrating gps and low cost ins. *PLANS 2004. Position Location and Navigation Symposium (IEEE Cat. No.04CH37556)*. doi: 10.1109/plans.2004.1308998.

[24] Yazhe Hu and Tomonari Furukawa. A high-resolution surface image capture and mapping system for public roads. *SAE International Journal of Passenger Cars - Electronic and Electrical Systems*, 10(2), 2017. doi: 10.4271/2017-01-0082.

[25] G Hunter, C Cox, and J Kremer. Development of a commercial laser scanning mobile mapping system - streetmapper. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.222.2382&rep=rep1&type=pdf.

[26] Anttoni Jaakkola, Juha Hyypp, Hannu Hyypp, and Antero Kukko. Retrieval algorithms for road surface modelling using laser-based mobile mapping. *Sensors*, 8(9):52385249, Jan 2008. doi: 10.3390/s8095238.

[27] Charles Johnson. Readiness of the road network for connected and autonomous vehicles. *Readiness of the road network for connected and autonomous vehicles*, Apr 2017.

[28] N. Karlsson, E. Di Bernardo, J. Ostrowski, L. Goncalves, P. Pirjanian, and M.e. Munich. The vslam algorithm for robust localization and mapping. *Proceedings of the 2005 IEEE International Conference on Robotics and Automation.* doi: 10.1109/robot.2005.1570091.

[29] Laurent Kneip, Margarita Chli, and Roland Siegwart. Robust real-time visual odometry with a single camera and an imu. *Procedings of the British Machine Vision Conference 2011*, 2011. doi: 10.5244/c.25.16.

[30] K. Konolige and M. Agrawal. Frameslam: From bundle adjustment to real-time visual mapping. *IEEE Transactions on Robotics*, 24(5):10661077, 2008. doi: 10.1109/tro.2008.2004832.

[31] Pramod Kumar, Pavel Ikonomov, Suren Dwivedi, Alamgir Choudhury, and Jorge Ro-

driguez. Laser scanning and modeling of a 3d road surface. *2008 Annual Conference & Exposition*, page 13.875.113.875.10, Jun 2008. URL https://peer.asee.org/4428.

[32] Piyathilaka Lasitha and R Munasinghe. Vision-only outdoor localization of two-wheel tractor for autonomous operation in agricultural fields. 08 2011.

[33] Wolfgang Lechner and Stefan Baumann. Global navigation satellite systems. *Computers and Electronics in Agriculture*, 25(1-2):6785, 2000. doi: 10.1016/s0168-1699(99)00056-3.

[34] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91110, 2004. doi: 10.1023/b:visi.0000029664.99615.94.

[35] Chenchi Luo and James H. Mcclellan. Robust geolocation estimation using adaptive ransac algorithm. *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010. doi: 10.1109/icassp.2010.5495817.

[36] R. Mehra. Approaches to adaptive filtering. *IEEE Transactions on Automatic Control*, 17(5):693698, 1972. doi: 10.1109/tac.1972.1100100.

[37] D. Nister, O. Naroditsky, and J. Bergen. Visual odometry. *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.* doi: 10.1109/cvpr.2004.1315094.

[38] Navid Nourani-Vatani and Paulo Vinicius Koerich Borges. Correlation-based visual odometry for ground vehicles. *Journal of Field Robotics*, 28(5):742768, Sep 2011. doi: 10.1002/rob.20407.

[39] W Ouyang and B Xu. Pavement cracking measurements using 3d laser-scan images. *Measurement Science and Technology*, 24(10):105204, 2013. doi: 10.1088/0957-0233/24/10/105204.

[40] Ignacio Parra, Miguel Angel Sotelo, David F. Llorca, C. Fernandez, A. Llamazares, N. Hernandez, and I. Garcia. Visual odometry and map fusion for gps navigation assistance. *2011 IEEE International Symposium on Industrial Electronics*, 2011. doi: 10.1109/isie.2011.5984266.

[41] Joern Rehder, Kamal Gupta, Stephen Nuske, and Sanjiv Singh. Global pose estimation with limited gps and long range visual odometry. *2012 IEEE International Conference on Robotics and Automation*, 2012. doi: 10.1109/icra.2012.6225277.

[42] J.z. Sasiadek and Q. Wang. Fuzzy adaptive kalman filtering for ins/gps data fusion and accurate positioning. *IFAC Proceedings Volumes*, 34(15):410415, 2001. doi: 10.1016/s1474-6670(17)40762-2.

[43] Helmut Heinrich Schmid. *Three-dimensional triangulation with satellites*. U.S. Gov. Print. Off.), 1975.

[44] Yun Shi, Shunping Ji, Zhongchao Shi, Yulin Duan, and Ryosuke Shibasaki. Gps-supported visual slam with a rigorous sensor model for a panoramic camera in outdoor environments. *Sensors*, 13(1):119136, 2012. doi: 10.3390/s130100119.

[45] Marcin Staniek. Stereo vision method application to road inspection. *The Baltic Journal of Road and Bridge Engineering*, 12(1):3847, 2017. doi: 10.3846/bjrbe.2017.05.

[46] J.-P. Tardif, Y. Pavlidis, and K. Daniilidis. Monocular visual odometry in urban environments using an omnidirectional camera. *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2008. doi: 10.1109/iros.2008.4651205.

[47] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org/, 2008.

[48] Kelvin C. P. Wang and Omar Smadi. Automated imaging technologies for pavement distress surveys. *TRANSPORTATION RESEARCH CIRCULAR*, Jul 2011. doi: 10.17226/22866.

[49] Lijun Wei, Cindy Cappelle, Yassine Ruichek, and Frdrick Zann. Intelligent vehicle localization in urban environments using ekf-based visual odometry and gps fusion. *IFAC Proceedings Volumes*, 44(1):1377613781, 2011. doi: 10.3182/20110828-6-it-1002.01965.

[50] Changchang Wu. Visualsfm : A visual structure from motion system. URL http://ccwu.me/vsfm/.

[51] Bisheng Yang, Lina Fang, and Jonathan Li. Semi-automated extraction and delineation of 3d roads of street scene from mobile laser scanning point clouds. *ISPRS Journal of Photogrammetry and Remote Sensing*, 79:8093, 2013. doi: 10.1016/j.isprsjprs.2013.01.016.

[52] Shichao Yang, Yu Song, Michael Kaess, and Sebastian Scherer. Pop-up slam: Semantic monocular plane slam for low-texture environments. *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016. doi: 10.1109/iros.2016.7759204.

[53] Si-Jie Yu, Sreenivas R. Sukumar, Andreas F. Koschan, David L. Page, and Mongi A. Abidi. 3d reconstruction of road surfaces using an integrated multi-sensory approach. *Optics and Lasers in Engineering*, 45(7):808818, 2007. doi: 10.1016/j.optlaseng.2006.12.007.

[54] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):13301334, 2000. doi: 10.1109/34.888718.

Appendices

Appendix A

2018 IEEE 88th Vehicular Technology Conference

AEKF-Based 3-D Localization of Road Surface Images with Sparse Low-Accuracy GPS Data

Diya Li
Department of Mechanical Engineering
Virginia Tech
Blacksburg VA, USA
lydia1@vt.edu

Yazhe Hu
Department of Mechanical Engineering
Virginia Tech
Blacksburg VA, USA
yazhehu@vt.edu

Tomonari Furukawa
Department of Mechanical Engineering
Virginia Tech
Blacksburg VA, USA
tomonari@vt.edu

***Abstract*—This paper presents a technique for localizing road surface images acquired by a downward-facing monocular camera on a vehicle with sparse low-accuracy Global Positioning System (GPS) readings. Images are collected by reading vehicle speed through on-board diagnostics (OBD) such that distance between two neighboring images is constant. The images are then stitched to create the road surface of arbitrary length. Lastly, the three-dimensional (3-D) road surface is created and globally corrected by using the GPS and the elevation map as well as the Adaptive Extended Kalman Filter (AEKF). The advantage of this technique is the possible deployment of a sparse low-accuracy GPS due to the use of the adaptive version of Extended Kalman Filter (EKF). The proposed technique was used for localization of local roads and highways of 6.9 km total length in Blacksburg, VA. The results of the localization show the reconstructed 3-D differ from the satellite imagery data only by 7.97%.**

***Index Terms*—adaptive estimation, extended Kalman filter, localization, sparse global position, visual odometry**

I. INTRODUCTION

Recent years have seen the growing need for collecting three-dimensional (3-D) data of public road surfaces. The 3-D road data affect the safety and the ride comfort of drivers and passengers significantly. The 3-D data are useful for not only effective road condition assessment and maintenance but also for future autonomous driving that needs to see the road condition in real time.

The past work on the road surface mapping can be classified into three categories. The first is using 3-D laser scanning which can produce high density point clouds of the road surface. Jaakkola, et al. [1] used a laser scanner and a Inertial Measurement Unit (IMU) for enhanced road surface mapping. It retrieves the road surface including paintings and kerbstones by the scanned point cloud data. Yang, et al. [2] employed a moving window filtering operation on the mobile laser scanning (MLS) data to classify the road surface from 3-D point cloud and further used the Global Positioning System (GPS) times to partition the points into different road sections. This method can accurately generate 3-D models of road surface, but it costs more than other survey methods and captures large amount of data that needs to be reduced in post-processing.

The second classification falls into the vision-based method. Image can provide more details of road surface and monocular image localization has been investigated intensively in the last few decades by a number of researchers. For vision-only deployment, scale drift is a noted issue for accurate localization, as errors are accumulated overtime at long distance travels [3]. One of the solutions for the drift error of the visual odometry is the Visual Simultaneous Localization and Mapping (SLAM) [3]–[5]. Tardif et al. [4] proposed the Mono-SLAM which localizes with 2.47% error over 2.5 km by decoupling the rotation and using epipolar constraints. Bundle Adjustment (BA) has also been used to resolve the drift error problem [6], [7], which is effective when features can be tracked over multiple frames and highly utilized in 3-D reconstruction. However, these methods are not robustly applied to low texture images which have few significant features. As limited salient features tracked, some works used dominant planar surface [8] to recover motion estimation for large-scale mapping. The aforementioned methods were used only with forward-facing cameras and were limited to apply in low texture environment.

The third category is the multi-sensor method. IMU [9], GPS [10], and Real-Time Kinematic GPS (RTK-GPS) [11] [12] are widely used sensors integrated with visual odometry for correcting the drift errors. Kneip et al. [9] validated that the use of IMU could mitigate the scale problem and recover relative camera motion [9]. Wei et al. [11] proposed the integration of vision and RTK-GPS with Extended Kalman Filter (EKF) to estimate the global trajectory and demonstrated its efficacy. Instead of using expensive high accuracy sensors, Agrawal and Konolige used a regular GPS receiver to provide global location information and landmarks to improve accuracy [10]. The past work can provide accurate global position, but the position update rate does not meet the requirement for high density road surface data.

This paper represents a vision-based road surface image 3-D localization technique with sparse low-accuracy GPS data and Adaptive Extended Kalman Filter (AEKF) model. By taking advantages of the geometry of the ground-facing camera system, the motion estimated by image shift between consecutive frames [13] [14] is used as prior in a recursive filter [15] to keep good estimation and avoid the triangulation error from 3-D point cloud. A GPS measurement combined with the elevation map is added to correct the accumulated position error from visual odometry in the global coordinate.

A B-spline curve method is proposed in this paper to solve sparse GPS data problem in high frame rate image pose correction. The proposed method reduces the translation error in long travel distance with adaptive noise covariance matrices, which are adjusted based on the residual and innovation error of recent measurements and states. Finally, the road map is generated by stitching high density road surface images with corrected 3-D image poses in global coordinate.

This paper is organized as follows. Sec. II demonstrates the fundamental of the proposed algorithm. Sec. III presents the approaches of estimating and correcting the image pose trajectory in world frame. Sec. IV shows the experiment system and analysis of the experimental results. The summary of this paper and the discussion of ongoing future work are presented in Sec. V.

II. 3-D Localization of Road Surface Images

A. 3-D Road Surface Mapping

The 3-D road surface map is generated by stitching road surface images with 3-D pose in the global frame. The problem is formulated by estimating the image 3-D pose, $\mathbf{x}_k$, in the world frame at time step k.

$$\mathbf{x}_k = f_k(\mathbf{x}_{k-1}) + \mathbf{v}_{k-1} \tag{1}$$

where $f_k(\cdot)$ is the state transition function and $\mathbf{v}_{k-1}$ is the process noise with zero mean and covariance $\mathbf{Q}_{k-1}$. The observation model for 3-D road surface image localization is defined with observation function $h_k(\cdot)$ which relates the current state and the observation $\mathbf{z}_k$:

$$\mathbf{z}_k = h_k(\mathbf{x}_k) + \mathbf{w}_k \tag{2}$$

where $\mathbf{w}_k$ is the observation noise with zero mean and noise covariance $\mathbf{W}_k$.

B. Extended Kalman Filter

The EKF can be a key framework of image localization, which gives an optimal estimation of the state from system models with noises and periodically updates the state from measurements. As descirbed above, the road surface image localization system can be formulated as the state equation, Eq. 1, and observation equation, Eq. 2. The conventional EKF framework consists of three basic stages: initialization, prediction and correction as shown in Fig. 1. Assuming random noise for all time steps and the initial state $\mathbf{x}_0$, the image pose can be determined in the prediction and correction stages.

1) Prediction: In the prediction stage, nonlinear state transition function, $f_k(\cdot)$, is used to estimate the current state of concern at time step k, $\widehat{\mathbf{x}}_k$, from its previous state, $\mathbf{x}_{k-1}$.

$$\widehat{\mathbf{x}}_k = f_k(\mathbf{x}_{k-1}) \tag{3}$$

$$\widehat{\mathbf{P}}_k = \mathbf{J}_{f_k}(\mathbf{x}_{k-1})\mathbf{P}_{k-1}{\mathbf{J}_{f_k}}^T(\mathbf{x}_{k-1}) + \mathbf{Q}_{k-1} \tag{4}$$

where $\widehat{\mathbf{x}}_k$ and $\widehat{\mathbf{P}}_k$ denote the priori state vector and error covariance matrix, $\mathbf{x}_k$ and $\mathbf{P}_k$ are the posteriori state vector and error covariance matrix, which is computed recursively, and $\mathbf{Q}_{k-1}$ is the covariance matrix of process noise.

2) Correction: In the correction stage, with having new observation or measurement $\mathbf{z}_k$, the image pose and error covariance of current state is updated:

$$\mathbf{x}_k = \widehat{\mathbf{x}}_k + \mathbf{K}_k(\mathbf{z}_k - \mathbf{H}_k\widehat{\mathbf{x}}_k) \tag{5}$$

$$\mathbf{P}_k = (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\widehat{\mathbf{P}}_k \tag{6}$$

where $\mathbf{H}_k$ is the Jacobian matrix of the observation function, $\mathbf{W}_k$ is the observation noise matrix and $\mathbf{K}_k$ is the Kalman gain.

$$\mathbf{K}_k = \widehat{\mathbf{P}}_k{\mathbf{H}_k}^T(\mathbf{H}_k\widehat{\mathbf{P}}_k{\mathbf{H}_k}^T + \mathbf{W}_k)^{-1} \tag{7}$$

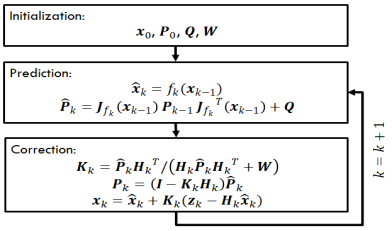


Fig. 1: The conventional framework of EKF

The conventional EKF model assumes that the process noise and observation noise are zero-mean Gaussian noise, thus it will work well for tuned model. However, if the error are not white noises, the filter will not perform as expected. One attempt to detect the filter divergence upon EKF is to add a fuzzy logic controller to detect the bias for preventing divergence [16]. The modified version of EKF is proposed which autonomously tunning process noise covariance to improve performance [17]. However, even though adaptively adjusting the noise covariances to avoid divergence, the conventional prediction model assumes smooth camera motion and constant velocities, which means unexpected error can occur during sudden changes in motion [15]. Also, comparing low rate measurement with required high rate state update for our application, the correction stage is not efficient.

III. Image Pose Trajectory

The proposed AEKF model of road surface image localization adaptively adjusts the noise covariance matrices as they highly affect the performance of EKF. The first part describes the overview of the AEKF model. Then, the principle of prediction is applied with visual odometry to enhance the certainty of local pose estimation. Finally, the image pose is globally corrected from interpolated GPS tag.

A. Overview

Unlike the conventional EKF, the proposed AEKF-based image localization uses visual odometry as priors to EKF for prediction and correct states by interpolated GPS tag with using adaptive covariance matrix.

Road surface image pose at time step k in global coordinate is represented as $\mathbf{x}_k$:

$$\mathbf{x}_k = [\mathbf{q}_k, \mathbf{r}_k]^T = [x_k, y_k, z_k, \alpha_k, \beta_k, \gamma_k]^T \tag{8}$$

where $\mathbf{q}_k = [x_k, y_k, z_k]$ is the image position in the global coordinate, in which x_k and y_k are the image position in east and north direction and z_k is in vertical direction. $\mathbf{r}_k = [\alpha_k, \beta_k, \gamma_k]$ is the orientation (Euler angle) with respect to the global coordinate.

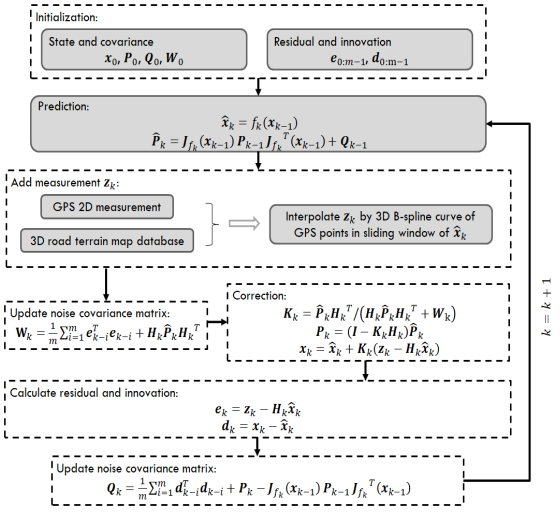


Fig. 2: Flowchart of the proposed AEKF-based 3-D image localization

Figure 2 shows the main structure of the proposed AEKF-based road surface image 3-D localization with adaptive covariance matrices of process and observation by the variation noise over the last m time steps. The novel parts in the proposed method are indicated with shadow boxes. The initial state $\mathbf{x}_0$ is estimated from GPS readings in the sliding window. Also, the first m time steps residual $\mathbf{e}_{0:m-1}$ and innovation $\mathbf{d}_{0:m-1}$ are initialized based on the fixed noise covariance, $\mathbf{W}_0$ and $\mathbf{Q}_0$.

In the prediction stage, the classical position and velocity model can work well in low dynamic scenario, but error will increase in sudden move. Thus, instead of using standard motion model with constant velocity, the proposed system uses visual odometry as priors.

$$\widehat{\mathbf{x}}_k = f_k(\mathbf{x}_{k-1}) = \begin{bmatrix} \widehat{\mathbf{q}}_k \\ \widehat{\mathbf{r}}_k \end{bmatrix} = \begin{bmatrix} \mathbf{q}_{k-1} + rotm(\mathbf{r}_{k-1})\widetilde{\mathbf{t}}_{k-1,k} \\ eul(\widetilde{\mathbf{R}}_{k-1,k} rotm(\mathbf{r}_{k-1})) \end{bmatrix} \tag{9}$$

where $\widetilde{\mathbf{R}}_{k-1,k}$ and $\widetilde{\mathbf{t}}_{k-1,k}$ are defined in Sec. III-B. And the Euler angles are converted to rotation matrix with function $rotm(\cdot)$ before calculating with the process model matrix and converted back using $eul(\cdot)$ function. $\mathbf{J}_{f_k}$ is the Jacobian matrix and $\mathbf{H}_k$ is an identity matrix here since the proposed method interpolates both position and rotation based on 3-D B-spline curve of GPS readings, as shown in Sec. III-C.

To update the noise covariance matrix, the proposed model calculates the residual $\mathbf{e}_k$ and innovation $\mathbf{d}_k$ and updates $\mathbf{W}_k$ and $\mathbf{Q}_k$ by residuals and innovations in the last m time steps. To avoid large noise at the beginning which may cause large error in pose estimation at the first few time steps, the proposed method initializes the residual and innovation without adaptive noise covariance in the first m time steps for updating noise covariance matrix. Also, the upper bounds of observation noise covariance is set to force visual odometry aligned with GPS curve so that the bad orientation from visual will not deviate from correct road too far in global coordinate.

B. Image Shift Estimation and Optimization

As the proposed system has downward-facing image capturing system, image shift is used to estimate motion. Based on the evaluated results of frame rate control, the translation in vehicle heading direction between consecutive frames is around 0.27 meters. Considering the shift between consecutive image is small, the proposed method assumes that the movement is in a 2D plane which has small change in elevation with respect to the previous frame. Therefore, N feature positions are detected and matched by the scale-invariant feature transform (SIFT) [20] and optimized motion, relative rotation $\mathbf{R}_{k-1,k}$, and translations $\mathbf{t}_{k-1,k}$ by least square rigid motion using singular value decomposition (SVD) [21].

$$(\mathbf{R}_{k-1,k}, \mathbf{t}_{k-1,k}) = \underset{\mathbf{R}\in\mathbb{R}^{2x2}, \mathbf{t}\in\mathbb{R}^2}{\operatorname{argmin}} \sum_{i=1}^{N} \left\| \mathbf{R}\mathbf{p}_{k-1}^i + \mathbf{t} - \mathbf{p}_k^i \right\|^2 \tag{10}$$

To estimate the optimal rotation, the matched feature point positions is centralized to remove the translation

$$\mathbf{R}_{k-1,k} = \underset{\mathbf{R}\in\mathbb{R}^{2x2}}{\operatorname{argmin}} \sum_{i=1}^{N} \left\| \mathbf{R}\mathbf{c}_{k-1}^i - \mathbf{c}_k^i \right\|^2 \tag{11}$$

where $\mathbf{c}_{k-1}^i$ and $\mathbf{c}_k^i$ are the centralized feature's position for image $\mathbf{I}_{k-1}$ and $\mathbf{I}_k$. Then with decomposing $\mathbf{C}_{k-1}^T \mathbf{C}_k$, the correlation matrix, by SVD, the optimal rotation is

$$\mathbf{R}_{k-1,k} = \mathbf{V}_k \mathbf{\Sigma}_k \mathbf{U}_k^T \tag{12}$$

Using the optimal rotation, the translation is

$$\mathbf{t}_{k-1,k} = \overline{\mathbf{p}}_k - \mathbf{R}_{k-1,k}\overline{\mathbf{p}}_{k-1} \tag{13}$$

where $\overline{\mathbf{p}}_{k-1}$ and $\overline{\mathbf{p}}_k$ are the center of feature position.

However, there will be outliers caused by the incorrectly matched features. Therefore, RANSAC is applied to remove the outliers to get better results. The outliers ratio is different from frame to frame as most of the outliers' are from shadows which vary in different driving directions. Thus, AEKF-based localization method is proposed to adaptively adjust the outliers ratio for each iteration in RANSAC with repeat times adjusted until convergence, which is called adaptive RANSAC (ARANSAC) [22].

For images collected from downward-facing cameras, moving shadows may cause wrong outliers removal, which results in a wrong motion estimation. To differentiate from incorrect matches, a threshold is applied over the inlier ratio in ARANSAC as shown in the above algorithm. Besides this, as the movement between two consecutive images is controlled

Algorithm 1 ARANASC with Motion Constraint

Initialization transformation $T_0 = (\mathbf{R}_0, \mathbf{t}_0)$
Get set of $inliers$ and $outliers$ with ARANSAC
if inliers ratio $I_0 < 70\%$ **then**
  Find T_{in}, I_{in} using $inliers$ with ARANSAC
  Find T_{out}, I_{out} using $outliers$ with ARANSAC
  if $\mathbf{t}_{x\,out} > 200$ **then**
    Finalize transformation $T_f = T_{out}$
  else
    Finalize transformation $T_f = T_{in}$
  end if
else
  Finalize transformation $T_f = T_0$
end if

based on the vehicle speed, the addition of motion constraint is included in ARANSAC by evaluating translation. Using this approach, the effects of mismatched features are negated from moving vehicle shadows.

Once the motion is optimized with centralized data, it is converting into 3-D model for the process model of AEKF:

$$\widetilde{\mathbf{R}}_{k-1,k} = \begin{bmatrix} \mathbf{R}_{k-1,k} & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix} \qquad \widetilde{\mathbf{t}}_{k-1,k} = s\mathbf{R}_{cw} \begin{bmatrix} \mathbf{t}_{k-1,k} \\ 0 \end{bmatrix} \quad (14)$$

where s is the ratio of pixel to meter calculated by calibrated fixed camera height and the camera's field of view (FOV). $\mathbf{R}_{cw}$ is the transformation from the camera coordinate to world coordinate.

The proposed process model can reduce the drift in sudden motion over time by keeping the overlapping region consistent between frames. This model also negates errors caused by moving vehicle shadow in motion estimation.

C. Interpolating Image Pose with Sparse Global Positions

To reduce the accumulated position and rotation errors in visual odometry, GPS data is used in the observation model for correcting image pose. Since the elevation data read from GPS is unreliable, the corresponding elevation from the elevation map [19] is collected to improve the global position accuracy. Sparse (low bandwidth 1 Hz) global position is insufficient for correcting the error in high image frame rate. By introducing a B-spline method, the corresponding GPS data for each image can be interpolated. To generate a B-spline curve, control points, GPS points $\mathbf{g} \in G$, is selected by a square slide window with width $2 * wid$ around estimated image position $\widehat{\mathbf{x}}_\mathbf{k}$. The size of slide window is also automatic adjusted based on the GPS data density in the window. GPS logs set $G = \{\mathbf{g}_0, \mathbf{g}_1, \cdots \mathbf{g}_n\}$ with $N = \{0, 1, \cdots n\}$ and the selected points are

$$\left\{ \mathbf{g}_{m_k^i} \mid m_k^i \in N \wedge \left(\widehat{x}_k - wid \le g_{x\,m_k^i} \le \widehat{x}_k + wid\right) \right. \\ \left. \wedge \left(\widehat{y}_k - wid \le g_{y\,m_k^i} \le \widehat{y}_k + wid\right) \right\} \quad (15)$$

And control point set consists of the $(n_{cp} + 1)$ selected GPS points: $\left\{ \mathbf{g}_{m_k^0}, \mathbf{g}_{m_k^1}, \cdots \mathbf{g}_{m_k^{n_{cp}}} \right\}$ with uniformly-spaced knot vector $\mathbf{t}_{kv}$ with $(n_{kt}+1)$ knots for d^{th} degree B-spline curve.

$$\mathbf{t}_{kv} = \begin{bmatrix} t_1 & t_1 & \cdots & t_{n_{kt}} \end{bmatrix} \quad (16)$$

The B-spline curve function is defined by a series of basis function, $B_{i,d}(u)$, in d degree:

$$B_{i,0}(u) = \begin{cases} 1, & t_i \le u \le t_{i+1} \\ 0, & \text{otherwise} \end{cases}$$
$$B_{i,d}(u) = \frac{u - t_i}{t_{i+d} - t_i} B_{i,d-1}(u) + \frac{t_{i+d+1} - u}{t_{i+d} - t_i} B_{i+1,d-1}(u) \quad (17)$$

And the line function $L_k(u)$ at time step k is

$$L_k(u) = \sum_{i=0}^{n_{cp}} \mathbf{g}_{m_k^i} B_{i,d}(u) \quad (18)$$

where uniformed $u \in [0, 1]$ and in our experiment,$d = 4$. To interpolate the corresponding image global position, we heavily sample the points in the range to find the closest point on the curve with respect to the position in $\widehat{\mathbf{x}}_\mathbf{k}$, which is $\widehat{\mathbf{q}}_\mathbf{k}$. The corresponding point on the curve is

$$j_k = \underset{j \in [0,1]}{\operatorname{argmin}}\, dist(L_k(j), \widehat{\mathbf{q}}_\mathbf{k}) \quad (19)$$

where function $dist$ calculates the least square distance between two points. Besides positions, the proposed method could also interpolate the orientation from the tangent vector $\mathbf{l}_{t_k}$ and binormal vector $\mathbf{l}_{b_k}$ derived from the first and second derivative of curve function.

$$\mathbf{l}_{t_k} = \frac{L'_k(j_k)}{\|L'_k(j_k)\|} = \begin{bmatrix} tx & ty & tz \end{bmatrix}^T \quad (20)$$

$$\mathbf{l}_{b_k} = \frac{L'_k(j_k) \times L''_k(j_k)}{\|L'_k(j_k) \times L''_k(j_k)\|} \quad (21)$$

As the z axis of vehicle frame should point up in the world frame, the sign of binormal vector is verified.

$$\mathbf{l}_{b_k} = -sign(\cos^{-1}(\mathbf{l}_{b_k}^T \cdot \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}) - \frac{\pi}{2}) * \mathbf{l}_{b_k} \quad (22)$$

As shown in Fig. 3, the orientation angles for each axis are found by the following equations:

$$\beta_{\mathbf{z}_k} = \sin^{-1}(tz), \mathbf{R}_{\beta_{\mathbf{z}_k}} = \begin{bmatrix} \cos(\beta_{\mathbf{z}_k}) & 0 & \sin(\beta_{\mathbf{z}_k}) \\ 0 & 1 & 0 \\ -\sin(\beta_{\mathbf{z}_k}) & 0 & \cos(\beta_{\mathbf{z}_k}) \end{bmatrix} \quad (23)$$

$$\gamma_{\mathbf{z}_k} = sign(ty) * \pi + \tan^{-1}(\frac{ty}{tx}) \quad (24)$$

The orientation along z axis is found by the angle between binormal vector and the z axis after the rotation along y axis, $\mathbf{r}_0 = \mathbf{R}_{\beta_{\mathbf{z}_k}} \cdot \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^T$.

$$\alpha_{\mathbf{z}_k} = sign\left((\mathbf{r}_0 \times \mathbf{l}_{b_k}) \cdot \mathbf{l}_{t_k}\right) \cos^{-1}\left(\mathbf{l}_{b_k}{}^T \cdot \mathbf{r}_0\right) \quad (25)$$

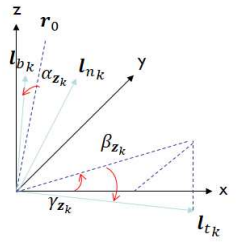


Fig. 3: Interpolate angles of axes from 3-D curve constructed by GPS points

Then the position estimate from the GPS for each image is realigned as:

$$\mathbf{z}_k = \begin{bmatrix} L_k(j)^T & \alpha_{\mathbf{z}_k} & \beta_{\mathbf{z}_k} & \gamma_{\mathbf{z}_k} \end{bmatrix}^T \in \mathbb{R}^6 \tag{26}$$

IV. EXPERIMENTS AND RESULTS

The experiment system consists of an camera system, a field programmable gate array (FPGA), a low-cost GPS, an on-board diagnostics (OBD) system and an on-board computer. The specifications and parameters used in the experiment are listed in Table. I. The image capturing system used in this paper, as shown in Fig. 4a, refers to Hu and Furukawa's design [18]. A pair of downward facing cameras are triggered and synchronized by a FPGA, and in our improved system, the system controls the image frame rate based on real time vehicle speed from OBD port which presents in most current vehicles, as shown in Fig. 4b. The system can maintain consistent overlapping region between consecutive images and reject the mismatched features from moving shadows by adding this motion constraint. This method efficiently covers road surface with enough images in different scenarios, such urban low speed, traffic light, highway, etc.

TABLE I: Experimental hardware specifications and parameters

Parameter	Value
Image sensor	Point Grey Flea3
FOV	$56°09' \times 43°36'$
Camera height: H	1.5202 m
Image resolution	1280×1024 (pixel)
GPS update rate	1Hz
GPS position accuracy	3-5m

This section will describe the details of the evaluation of frame rate control, IV-A, and localization, IV-B. The data is collected in both urban and highway environment to test the performance of the frame rate control in dynamic speed. To validate the proposed AEKF-based method of localization, images are reprojected to global frame to compare with satellite images.

A. Frame Rate Control

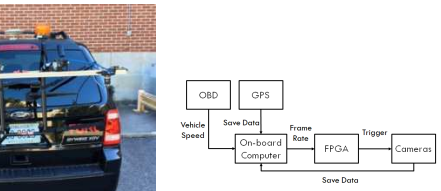
The height, H, of the image plane from the road surface is calibrated. Knowing the camera field of view along the vehicle moving direction, $a = 43°36'$, from Table. I, the length of image along the vehicle moving direction L is estimated

$$L = 2H * tan(\frac{a}{2}) \tag{27}$$

To keep enough overlapping region for good features matching, the frame rate fr is defined by

$$fr = \frac{Vn}{3.6L} \tag{28}$$

where V is the vehicle speed in km/h and n is the number of image that cover the same region in L.



(a) Hardware platform (b) System layout

Fig. 4: System overview

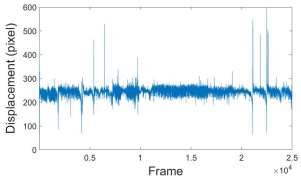


Fig. 5: Dynamically frame rate control based on vehicle speed and the evaluation of overlapping

The translation between I_k and I_{k-1} in pixels with the proposed method is calculated to evaluate the performance of frame rate control. In results, as shown in Fig.5, the average translation with respect to vehicle moving direction is about 235 pixels during driving which is around quarter of image length along the vehicle moving direction as designed. The lower bound of pixel displacement is around 200 which used as a threshold of motion constraint in Alg.1. In Fig.5, there are some peaks and valleys in the plot, these happen when the vehicle brakes or accelerates in a short time.

B. Localization Results

One of the experiment paths consists of both urban, which has more sharp turning curves, and highway, which has smooth straight lines, located in Blacksburg, Virginia, as shown in Fig. 6.

GPS data is inaccurate locally but does not suffer from accumulation of errors. In long range, accumulation of errors of visual odometry results in the end position being far off the destination. The corrected trajectory with our estimation method is shown in Fig. 7. The result shows that the visual

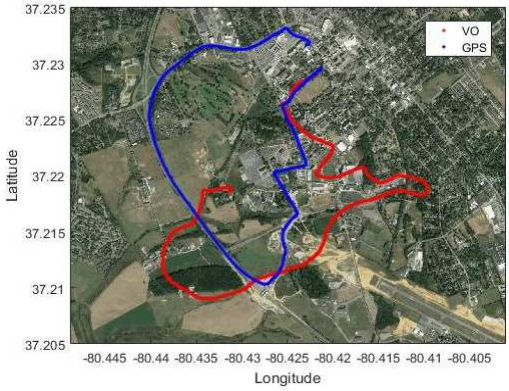


Fig. 6: Comparison between the trajectory estimated by vision-only and raw GPS in world frame

odometry has been corrected along with GPS position. Also, it is obvious to find that errors around turns are more than straight roads, and accumulated error caused drift in long range. Some positions have large drift since the orientation of image is not corrected efficiently, which leading to accumulation of errors, but they will be draw back from position measurements.

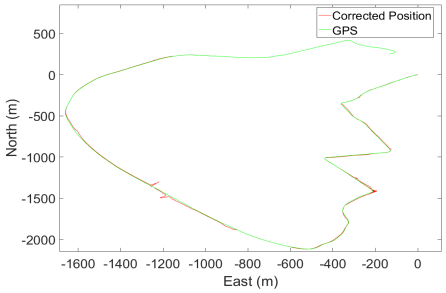


Fig. 7: Corrected image positions and raw GPS data

By projecting images back to satellite imagery and comparing the position of road markings, as shown in Fig. 8, the localization accuracy is validated by translation error over travel distance. It shows a part of result that the high resolution road surface images are stitched over satellite image with estimated global positions and orientations. Red points indicate the corresponding position on the satellite image. Also, the pedestrian walkway and the edge of road marking on the road surface images are closely matched with the satellite image.

The total frames in this experiment is 25383, and the route starts and ends in urban area. The translation error is found by tracking road markings in the image sequences, and most of comparison points are in the urban area. From the results listed in Table II, the translation error increases over time. More errors occur on sharp turns of urban road, while less errors accumulated on highways. The best result of global

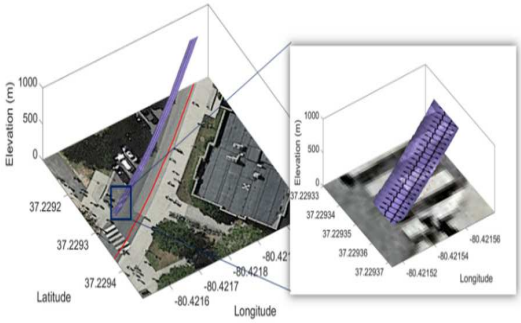


Fig. 8: Images projected over the satellite image

position estimation obtained is in shorter distance, 1.157 km, with translation error 0.85%. For the longest distance, 6.9 km, it has the translation error 7.97%.

TABLE II: Compare translation error over travel distance

Distance (m)	Frames	Translation error %
491.3	1864	1.70
1156.9	4333	0.85
1478.2	5517	5.95
5836.0	22782	6.85
6902.9	25227	7.97

V. CONCLUSION

This paper has presented an AEKF-based 3-D road surface image localization method with sparse low-accuracy GPS. The proposed technique improves the position estimation by using image shift as priors and adding motion constraint to remove moving shadow effects. Owing to the sparse GPS data, a B-Spline curve method interpolates the 3-D pose of each frame from GPS readings and the elevation map. Since the noise covariances are adaptively adjusted based on recent residuals and innovations, the proposed technique can reduce the translation error and generate road surface map in 3-D.

The evaluation of position accuracy was investigated by the translation error at the end position over travel distance. The experiment localized all images over a long range, 6.9 km, with translation error of 7.97%. The best result was in shorter distances, 1.157 km, with translation error of 0.85%. The results have demonstrated the effectiveness and applicability of the approach. As the error will increase in sharp turns, in the future work, the proposed method can be extended by aligning road marking positions in road surface images with their position in satellite images around the estimated trajectory. In addition, more accurate inertial sensors can be added for improving orientation estimation and evaluation.

REFERENCES

[1] Jaakkola, Anttoni, et al. “Retrieval Algorithms for Road Surface Modelling Using Laser-Based Mobile Mapping.” Sensors, vol. 8, no. 9, Jan. 2008, pp. 5238–5249., doi:10.3390/s8095238.

[2] Yang, Bisheng, et al. "Semi-Automated Extraction and Delineation of 3D Roads of Street Scene from Mobile Laser Scanning Point Clouds." ISPRS Journal of Photogrammetry and Remote Sensing, vol. 79, 2013, pp. 80–93., doi:10.1016/j.isprsjprs.2013.01.016.
[3] N. Karlsson, E. D. Bernardo, J. Ostrowski, L. Goncalves, P. Pirjanian, and M. Munich, "The vSLAM Algorithm for Robust Localization and Mapping," Proceedings of the 2005 IEEE International Conference on Robotics and Automation.
[4] J.-P. Tardif, Y. Pavlidis, and K. Daniilidis, "Monocular Visual Odometry in Urban Environments Using an Omnidirectional Camera," 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2008.
[5] H. Choi, D. Kim, J. Hwang, C. Park, and E. Kim, "Efficient Simultaneous Localization and Mapping Based on Ceiling-View: Ceiling Boundary Feature Map Approach," Advanced Robotics, vol. 26, no. 5-6, pp. 653–671, 2012.
[6] D. P. Frost, O. Kahler, and D. W. Murray, "Object-Aware Bundle Adjustment for Correcting Monocular Scale Drift," 2016 IEEE International Conference on Robotics and Automation (ICRA), 2016.
[7] K. Konolige and M. Agrawal, "FrameSLAM: From Bundle Adjustment to Real-Time Visual Mapping," IEEE Transactions on Robotics, vol. 24, no. 5, pp. 1066–1077, 2008.
[8] R. Siddiqui and S. Khatibi, "Robust Visual Odometry Estimation of Road Vehicle from Dominant Surfaces for Large-Scale Mapping," IET Intelligent Transport Systems, vol. 9, no. 3, pp. 314–322, Jan. 2015.
[9] L. Kneip, M. Chli, and R. Siegwart, "Robust Real-Time Visual Odometry with a Single Camera and an IMU," Procedings of the British Machine Vision Conference 2011, 2011.
[10] M. Agrawal and K. Konolige, "Real-time Localization in Outdoor Environments using Stereo Vision and Inexpensive GPS," 18th International Conference on Pattern Recognition (ICPR06), 2006.
[11] L. Wei, C. Cappelle, Y. Ruichek, and F. Zann, "Intelligent Vehicle Localization in Urban Environments Using EKF-based Visual Odometry and GPS Fusion," IFAC Proceedings Volumes, vol. 44, no. 1, pp. 13776–13781, 2011.
[12] J. Rehder, K. Gupta, S. Nuske, and S. Singh, "Global Pose Estimation with Limited GPS and Long Range Visual Odometry," 2012 IEEE International Conference on Robotics and Automation, 2012.
[13] N. Nourani-Vatani and P. V. K. Borges, "Correlation-Based Visual Odometry for Ground Vehicles," Journal of Field Robotics, vol. 28, no. 5, pp. 742–768, Sep. 2011.
[14] M. O. A. Aqel, M. H. Marhaban, M. I. Saripan, N. B. Ismail, and A. Khmag, "Optimal Configuration of a Downward-Facing Monocular Camera for Visual Odometry," Indian Journal of Science and Technology, vol. 8, no. 32, Oct. 2016.
[15] P. F. Alcantarilla, L. M. Bergasa, and F. Dellaert, "Visual Odometry Priors for Robust EKF-SLAM," 2010 IEEE International Conference on Robotics and Automation, 2010.
[16] Sasiadek, J.z., and Q. Wang. "Fuzzy Adaptive Kalman Filtering for INS/GPS Data Fusion and Accurate Positioning." IFAC Proceedings Volumes, vol. 34, no. 15, 2001, pp. 410–415., doi:10.1016/s1474-6670(17)40762-2.
[17] Ding, Weidong, et al. "Improving Adaptive Kalman Estimation in GPS/INS Integration." Journal of Navigation, vol. 60, no. 03, Sept. 2007, p. 517., doi:10.1017/s0373463307004316.
[18] Y. Hu and T. Furukawa, "A High-Resolution Surface Image Capture and Mapping System for Public Roads," SAE International Journal of Passenger Cars - Electronic and Electrical Systems, vol. 10, no. 2, 2017.
[19] K. L. El-Ashmawy, "Investigation of the Accuracy of Google Earth Elevation Data," Artificial Satellites, vol. 51, no. 3, Jan. 2016.
[20] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," International Journal of Computer Vision, vol. 60, no. 2, pp. 91–110, 2004.
[21] D. Eggert, A. Lorusso, and R. Fisher, "Estimating 3-D Rigid Body Transformations: a Comparison of Four Major Algorithms," Machine Vision and Applications, vol. 9, no. 5-6, pp. 272–290, Jan. 1997.
[22] C. Luo and J. H. Mcclellan, "Robust Geolocation Estimation Using Adaptive RANSAC Algorithm," 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, 2010.